TARGETED REDUCTION OF CAUSAL MODELS

Armin Kekić **Bernhard Schölkopf** **Michel Besserve**
Max Planck Institute for Intelligent Systems, Tübingen, Germany
{armin.kekic,bs,besserve}@tue.mpg.de

ABSTRACT

Why does a phenomenon occur? Addressing this question is central to most scientific inquiries and often relies on simulations based on differential equations. As scientific models become more intricate, deciphering the causes behind phenomena in high-dimensional spaces of interconnected variables becomes increasingly challenging. We introduce *Targeted Causal Reduction* (TCR), a method for condensing complex intervenable models into a concise set of causal factors that explain a specific target phenomenon. We propose an information theoretic objective to learn TCR from interventional data of simulations, establish identifiability for continuous variables under shift interventions and present a practical algorithm for learning TCRs. Its ability to generate interpretable high-level explanations from complex models is demonstrated on toy and mechanical systems, illustrating its potential to assist scientists in the study of complex phenomena in a broad range of disciplines.

1 INTRODUCTION

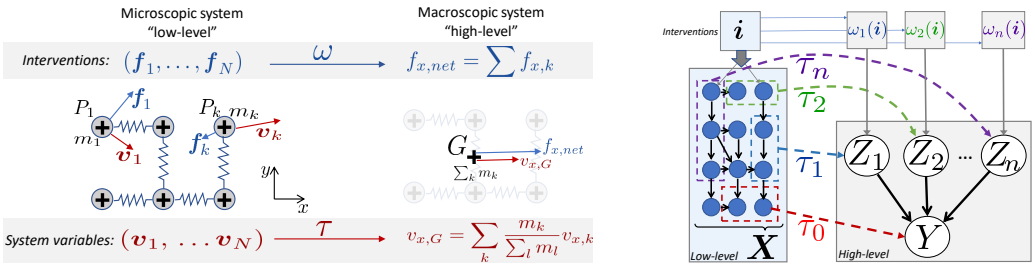
Numerical models are indispensable in science for simulating real-world systems and generating *etiological explanations*—identifying the causes of specific phenomena. General circulation models, for example, shed light on the causes of global warming (Grassl, 2000), while computational brain models explore the origins of neurological pathologies (Breakspear, 2017; Deco & Kringelbach, 2014). These examples illustrate the increasing complexity of numerical scientific models, designed to capture the large number of mechanisms at play in these systems. However, this complexity impacts the ability to generate high-level explanations, understandable by scientists and decision makers (Reichstein et al., 2019; Safavi et al., 2023).

Consider a system of point masses connected by springs, shown in Fig. 1a, where each mass is influenced by random external forces. Its trajectory can be accurately predicted by simulating the coupled equations of motion of the individual masses. However, if we are only interested in a particular macroscopic "target" variable of this system: the horizontal speed of the system's center of mass, classical physics tells us that its motion depends only on the sum of all horizontal components of external forces applied over time. We thus obtain a form of Causal Model Reduction (CMR): a much simpler model that accurately accounts for the effect of interventions in the system on the target variable.

This highlights the core elements needed for CMR in scientific models: (1) there is a defined macroscopic target variable, (2) continuous low-level variables are reduced to a smaller set of continuous high-level variables, and (3) low-level interventions are soft: exerted forces modify the future trajectory but to not suppress the influence of other factors such as the past state of the system.

In this paper, we introduce *Targeted Causal Reduction* (TCR), depicted in Fig. 1b, a novel approach designed to simplify complex simulations, viewed as *low-level models*, into *high-level models*, focused on explaining causal influences on an observable target variable Y. The key signal we use for learning are *interventions* applied to the low-level variables, which are mapped to high-level interventions in a way that captures the causal influences on Y in a concise and interpretable way. We formulate this learning objective as a KL divergence between the fitted high-level interventional model and the reduced low-level interventional distribution, leading to a practical learning algorithm. Applications to scientific simulations based on differential equations demonstrate the accuracy and interpretability of our approach.[1]

[1]We refer to the appendix for more background and related work (A).



(a) (b)

Figure 1: **Targeted Causal Reduction.** (a) Example targeted model reduction: a model of the dynamics of a system of point masses connected by springs can be reduced to the trajectory of its center of gravity. (b) Overview of TCR. Low-level variables $\boldsymbol{X}$ (simulation) are mapped to high-level variables $(\boldsymbol{Z}, Y)$ with a fixed causal structure. The target Y is known, while the causes $\boldsymbol{Z}$ and the high-level causal mechanism are learned. Additionally, we learn a mapping from low-level shift interventions $\boldsymbol{i}$ to high-level shift interventions $\boldsymbol{\omega}(\boldsymbol{i})$.

2 Causal Models

Causal dependencies between variables can be described using *Structural Causal Models* (SCM).

Definition 2.1 (SCM (Peters et al., 2017))**.** *An n-dimensional structural causal model is a triplet* $\mathcal{M} = (\mathcal{G}, \mathbb{S}, P_{\boldsymbol{U}})$ *consisting of: (i) a joint distribution* $P_{\boldsymbol{U}}$ *over exogenous variables* $\{U_j\}_{j \leq n}$*, (ii) a directed graph* $\mathcal{G}$ *with n vertices, (iii) a set* $\mathbb{S} = \{X_j := f_j(\boldsymbol{Pa}_j, U_j), j = 1, \ldots, n\}$ *of structural equations, where* $\boldsymbol{Pa}_j$ *are the variables indexed by the set of parents of vertex* j *in* $\mathcal{G}$*, such that for almost every* $\boldsymbol{u}$*, the system* $\{x_j := f_j(\boldsymbol{pa}_j, u_j)\}$ *has a unique solution* $\boldsymbol{x} = \boldsymbol{g}(\boldsymbol{u})$.

Soft interventions in SCMs involve replacing one or more structural equations while keeping causal dependencies in the original graph. It transforms the model $\mathcal{M}$ into an intervened model $\mathcal{M}^{(\boldsymbol{i})} = (\mathcal{G}, \mathbb{S}^{(\boldsymbol{i})}, P_{\boldsymbol{U}}^{(\boldsymbol{i})})$, where $\boldsymbol{i}$ is the vector parameterizing the intervention. The base probability distribution of the unintervened model is denoted $P^{(0)}(\boldsymbol{X})$ and the interventional distribution is denoted $P^{(\boldsymbol{i})}(\boldsymbol{X})$. Here, we focus on *shift interventions* that modify the structural equation of endogenous variable l through shifting it by a scalar parameter i, *i.e.* $\{X_l := f_l(\mathbf{Pa}_l, U_l)\} \mapsto \{X_l := f_l(\mathbf{Pa}_l, U_l) + i\}$.

3 TCR Framework

Simulations based on the numerical solution of differential equations can often be expressed as SCMs; This notably includes Ordinary (ODE) (Mooij et al., 2013) and Stochastic Differential Equations (SDE) (Hansen & Sokol, 2014). The core idea behind TCR is viewing the simulation as a low-level causal model from which we can get (un-)intervened samples. We then learn a mapping to a smaller high-level causal model. As a learning objective, we use *consistency* between the two causal models, *i.e.* the following two paths should be (approximately) equivalent: (i) first intervening on the low-level model and then mapping to high-level variables, or (ii) first mapping to high-level variables and then intervening in the high-level model. Our framework has the following elements, depicted in Fig. 1b:

(1) A low-level SCM $\mathcal{L}$ with N endogenous variables $\{X_1, X_2, \ldots, X_N\}$ and corresponding exogenous variables $\{U_k\}_{k=1..N}$ equipped with joint distribution $P(\boldsymbol{U})$. A set of low-level interventions parameterized by vector $\boldsymbol{i} \in \mathcal{I}$ with distribution $P(\boldsymbol{i})$, with each component i_k affecting a unique endogenous variable X_k.

(2) A class of high-level SCMs $\{\mathcal{H}_{\boldsymbol{\gamma}}\}_{\gamma \in \Gamma}$ with (n+1) endogenous variables $\{Y, Z_1, \ldots, Z_n\}$ and associated exogenous variables $\{R_k\}_{k=0..n}$, equipped with a factorized distribution $P(\boldsymbol{R}) = \prod P_{R_k}$. A set of high-level interventions parametrized by vector $\boldsymbol{j} \in \mathcal{J}$, with each component j_k affecting a single node Z_k. In contrast to the (fixed) low-level model, the high-level model parameters $\boldsymbol{\gamma}$ are learned.

These two models are linked by a transformation with two surjective maps τ and ω from low- to high-level endogenous variables and interventions, respectively, which decompose as

$$\tau = (\tau_0, \tau_1, \tau_2, \ldots, \tau_n) \text{ with } \tau_k : x \mapsto \bar{\tau}_k(x_{\pi(k)}) \tag{1}$$

$$\omega = (\omega_0, \omega_1, \omega_2, \ldots, \omega_n) \text{ with } \omega_k : \boldsymbol{i} \mapsto \bar{\omega}_k(\boldsymbol{i}_{\pi(k)}) \tag{2}$$

where π is a so-called alignment function from $[0\,..\,n]$ to non-overlapping subsets of $[1\,..\,N]$. Importantly, $\tau_0(\boldsymbol{X}) = Y$ maps to the target variable which encodes a phenomenon of interest; it is assumed fixed and known.[2]

The high-level model involves the following mechanisms which need to be learned: (1) The marginal distribution of each high-level cause $P^{(\boldsymbol{j})}(Z_k)$ in all high-level interventional settings $\boldsymbol{j}$. (2) The mechanism $P(Y|\boldsymbol{Z})$ mapping high-level causes to Y, comprised of the distribution of the exogenous variable R_0 and the map $(Z_1, ..., Z_n, R_0) \mapsto f_\gamma(Z_1, ..., Z_n, R_0) =: Y$.

Tau map. For interpretability, we assume a linear τ-map, represented as a vector $\boldsymbol{\tau}$ such that:

$$\boldsymbol{X} \mapsto \begin{bmatrix} Y \\ \boldsymbol{Z} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\tau}_0^\top \\ \vdots, \\ \boldsymbol{\tau}_n^\top \end{bmatrix} \boldsymbol{X} = \left[\bar{\boldsymbol{\tau}}_0^\top \boldsymbol{X}_{\pi(0)}\,, \quad \ldots \bar{\boldsymbol{\tau}}_n^\top \boldsymbol{X}_{\pi(n)} \right]^\top .$$

Omega map. We focus on *shift interventions* and map the vector $\boldsymbol{i}$ of low-level interventions on the nodes in $\pi(k)$ to a scalar shift intervention on the mechanism of each Z_k. We assume each map ω_k to be linear with vector $\boldsymbol{\omega}_k$ such that $\omega_k(\boldsymbol{i}) = \boldsymbol{\omega}_k^\top \boldsymbol{i} = \bar{\boldsymbol{\omega}}_k^\top \boldsymbol{i}_{\pi(k)}$.

Causal consistency loss. It is not always possible to achieve a transformation that guarantees consistency of low- and high-level models for almost all interventions. As a consequence, we allow for the consistency between models to be approximate. To ensure that this approximation is as accurate as possible, we minimize the expected KL divergence between the pushforward by the transformation τ of the low-level interventional distributions that we denote $\widehat{P}_\tau^{(\boldsymbol{i})}(Y, \boldsymbol{Z}) = \tau_\#[P_{\mathcal{L}}^{(\boldsymbol{i})}(\boldsymbol{X})]$, and the corresponding interventional distribution of the high-level model $P^{(\omega(\boldsymbol{i}))}$, leading to the consistency loss

$$\mathcal{L}_{cons} = \mathbb{E}_{\boldsymbol{i}\sim P(\boldsymbol{i})} \left[KL\left(\widehat{P}_\tau^{(\boldsymbol{i})}(Y, \boldsymbol{Z}) \| P^{(\omega(\boldsymbol{i}))}(Y, \boldsymbol{Z}) \right) \right] . \tag{3}$$

Since the KL divergence is challenging to learn non-parametrically, we make a Gaussian assumption on the densities. This allows us to obtain an analytic expression for the loss based on second order statistics (see expression in App. F.1).

Regularization. To encourage the high-level variables to focus on different parts of the input, we regularize the overlap between τ- and ω maps, as a differentiable way to encode the alignment π (see Eq. (12) and (13) in App.C). The details of the learning procedure are described in Algorithm 1 in App. C.

Theoretical analysis. If we further assume that the low-level model is linear Gaussian, we can find an analytical solution for TCR which provides intuition about the reductions produced by our framework and how many interventions are necessary to learn it. The full identifiability analysis and discussion of the properties of the consistency loss (3) are presented in App. B.

4 Experiment: Spring-Mass System

We simulate a two-dimensional system of four point masses with different weights connected by springs to their respective nearest neighbors, similar to the motivating example in Sec. 1. Initially, the masses are arranged in a rectangle in space such that the springs are at rest length. The masses have a random initial velocity, as shown in Fig. 2(a). As interventions, we apply random shifts to the velocities in x- and y-direction of each mass. The target of the simulation is the center of mass speed in $(1, 1)$-direction. We learn a TCR with two high-level causes. The full experimental details are given in App. G.4.

While the velocities of the individual masses are coupled, the center of mass velocities in x- and y-direction of the system as a whole are independent, since the system is freely moving in space. The

[2] Additionally, ω_0 is assumed to be a trivial constant map $\boldsymbol{i} \to 0$, to ensure that the high-level target variable cannot be directly intervened upon, as we want to explain the changes in Y exclusively through changes of its high-level causes.

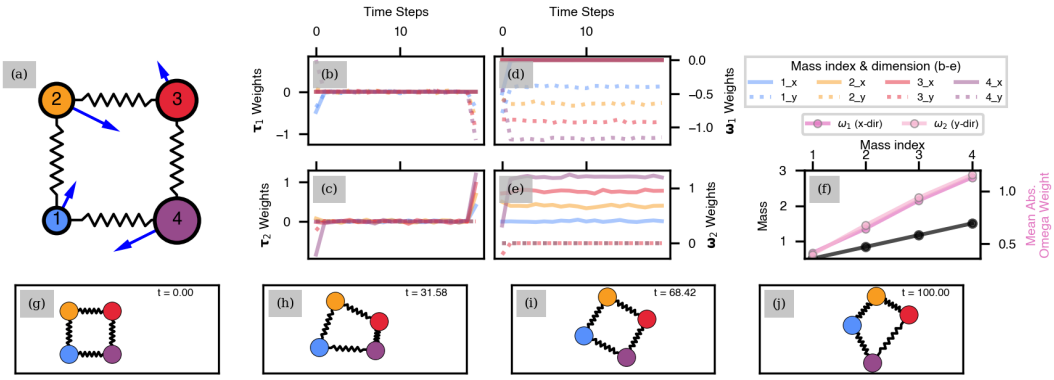


Figure 2: **Spring-mass system experiment.** (a) Simulated system of four point masses with different weights connected by springs and with random initial velocity (blue arrows). The target of the simulation is the center of mass speed in $(1, 1)$-direction. (b-e) Learned $\boldsymbol{\tau}$- and $\boldsymbol{\omega}$-weights corresponding to velocity components in x- and y-direction for a TCR with two high-level variables. The learned high-level mechanism is $f(\boldsymbol{Z}) \approx -0.226Z_1 + 0.220Z_2$. (f) Comparison between masses and learned omega weights. For the first high-level variable the mean omega weights corresponding to the x-direction are shown and for the second variable those for the y-direction. (g-j) Example trajectory for an unintervened system.

learned TCR correctly identifies these as the two independent causes of the target, with variable Z_1 corresponding to the y-direction and Z_2 to the x-direction. On average, each mass receives a similar shift in velocity through the applied interventions. However, since the masses are different, the shifts correspond to different contributions to the momentum of the system as a whole impacting the target. This is reflected in the relative weights of the learned maps being proportional to the weight of each point mass, as shown in Fig 2(f).

In App. G.1 we show additional experiments for synthetic linear Gaussian low-level models to validate the learning algorithm and gain intuition on the found reductions. We also demonstrate TCR on an ODE simulation of a ball moving in a double well potential and another configuration of the spring-mass system with two groups of interconnected masses.

5 Discussion

We introduce a novel approach for understanding complex simulations by learning high-level causal explanations from low-level models. Our Targeted Causal Reduction (TCR) framework leverages interventions to obtain simplified, high-level representations of the causes of a target phenomenon. We formulate the intervention-based consistency constraint as an information theoretic learning objective, which favors the most causally relevant explanations of the target. Under linearity and Gaussianity assumptions, we provide analytical solutions and study their uniqueness, which provides insights into TCR's governing principles. We provide an algorithm for linear TCR and show it can effectively uncover the key causal factors influencing a phenomenon of interest. We demonstrate TCR on both synthetic models and scientific simulations, highlighting its potential for addressing the challenges posed by increasingly complex systems in scientific research.

While we develop a CMR framework to learn high-level explanations for simulations, the simulation itself does not have to be explicitly formulated as a causal model and the causal relationships between variables in $\boldsymbol{X}$ do not have to be known a priori. The only additional element needed to learn TCR is a notion of shift-interventions. We think that most scientific simulations based on differential equations naturally allow for a reasonable notion of shift interventions.

Limitations and future work. To foster interpretability and tractability, we made Gaussian approximations and used linear $\boldsymbol{\tau}$ and $\boldsymbol{\omega}$ maps. While this has clear benefits, this may be too limiting for some complex simulations.Additionally, our method relies on performing a large number of interventions in simulation runs, which represents an additional cost in the context of large-scale simulation. How to make the algorithm scale to this setting is left to future work.

REFERENCES

Brandon Amos, Ivan Dario Jimenez Rodriguez, Jacob Sacks, Byron Boots, and J Zico Kolter. Differentiable MPC for end-to-end planning and control. *arXiv preprint arXiv:1810.13400*, 2018. 10

Sander Beckers and Joseph Y Halpern. Abstracting causal models. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pp. 2678–2685, 2019. 8, 10

Sander Beckers, Frederick Eberhardt, and Joseph Y Halpern. Approximate causal abstractions. In *Uncertainty in artificial intelligence*, pp. 606–615. PMLR, 2020. 10, 12

Michel Besserve and Bernhard Schölkopf. Learning soft interventions in complex equilibrium systems. In *Uncertainty in Artificial Intelligence*, pp. 170–180. PMLR, 2022. 7

Michel Besserve, Arash Mehrjou, Rémy Sun, and Bernhard Schölkopf. Counterfactuals uncover the modular structure of deep generative models. In *ICLR2020*, 2020. 10

BN Biswas, Somnath Chatterjee, SP Mukherjee, and Subhradeep Pal. A discussion on euler method: A review. *Electronic Journal of Mathematical Analysis and Applications*, 1(2):2090–2792, 2013. 8

Stephan Bongers, Patrick Forré, Jonas Peters, Bernhard Schölkopf, Joris M Mooij, et al. Foundations of structural causal models with cycles and latent variables. *arXiv preprint arXiv:1611.06221*, 2016. 7

Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021. 9

Michael Breakspear. Dynamic models of large-scale brain activity. *Nature neuroscience*, 20(3):340–352, 2017. 1

Kailash Budhathoki, Lenon Minorics, Patrick Blöbaum, and Dominik Janzing. Causal structure-based root cause analysis of outliers. In *International Conference on Machine Learning*, pp. 2357–2369. PMLR, 2022. 10

Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning. *arXiv preprint arXiv:1412.2309*, 2014. 7, 10

Krzysztof Chalupka, Tobias Bischoff, Pietro Perona, and Frederick Eberhardt. Unsupervised discovery of el nino using causal feature learning on microlevel climate data. *arXiv preprint arXiv:1605.09370*, 2016. 7, 10

Gustavo Deco and Morten L Kringelbach. Great expectations: using whole-brain computational connectomics for understanding neuropsychiatric disorders. *Neuron*, 84(5):892–905, 2014. 1

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586, 2021. 10

Atticus Geiger, Chris Potts, and Thomas Icard. Causal abstraction for faithful model interpretation. *arXiv preprint arXiv:2301.04709*, 2023a. 8, 10

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D Goodman. Finding alignments between interpretable causal variables and distributed neural representations. *arXiv preprint arXiv:2303.02536*, 2023b. 10

Philipp Geiger and Christoph-Nikolas Straehle. Learning game-theoretic models of multiagent trajectories using implicit layers. *arXiv preprint arXiv:2008.07303*, 2020. 10

Hartmut Grassl. Status and improvements of coupled general circulation models. *Science*, 288(5473):1991–1997, 2000. 1

Niels Hansen and Alexander Sokol. Causal interpretation of stochastic differential equations. *Electronic Journal of Probability*, 19:1 – 24, 2014. 2

Wendong Liang, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal component analysis. *arXiv preprint arXiv:2305.17225*, 2023. 10

Riccardo Massidda, Atticus Geiger, Thomas Icard, and Davide Bacciu. Causal abstraction with soft interventions. In *2nd Conference on Causal Learning and Reasoning*, 2023. 10

Annie Millet and Pierre-Luc Morien. On implicit and explicit discretization schemes for parabolic spdes in any dimension. *Stochastic Processes and their Applications*, 115(7):1073–1106, 2005. 9

Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. From ordinary differential equations to structural causal models: the deterministic case. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, pp. 440–448, Arlington, Virginia, USA, 2013. AUAI Press. 2

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference – Foundations and Learning Algorithms*. MIT Press, 2017. 2, 7

Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and fnm Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566 (7743):195–204, 2019. 1

Eigil F Rischel and Sebastian Weichwald. Compositional abstraction error and a category of causal models. In *Uncertainty in Artificial Intelligence*, pp. 1013–1023. PMLR, 2021. 10, 12

Paul K Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. *arXiv preprint arXiv:1707.00819*, 2017. 8, 10

Shervin Safavi, Theofanis I Panagiotaropoulos, Vishal Kapoor, Juan F Ramirez-Villegas, Nikos K Logothetis, and Michel Besserve. Uncovering the organization of neural circuits with generalized phase locking analysis. *PLOS Computational Biology*, 19(4):e1010983, 2023. 1

Timothy Sauer. Computational solution of stochastic differential equations. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(5):362–371, 2013. 8

Chandler Squires, Anna Seigal, Salil Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *40th International Conference on Machine Learning*, 2023. 10

Julius von Kügelgen, Michel Besserve, Wendong Liang, Luigi Gresele, Armin Kekić, Elias Bareinboim, David M Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *arXiv preprint arXiv:2306.00542*, 2023. 10

Julius Von Kügelgen, Abdirisak Mohamed, and Sander Beckers. Backtracking counterfactuals. In *Conference on Causal Learning and Reasoning*, pp. 177–196. PMLR, 2023. 10

Fabio Massimo Zennaro, Máté Drávucz, Geanina Apachitei, W Dhammika Widanage, and Theodoros Damoulas. Jointly learning consistent causal abstractions over multiple interventional distributions. *arXiv preprint arXiv:2301.05893*, 2023. 10, 12

Yuchen Zhu, Kailash Budhathoki, Jonas Kuebler, and Dominik Janzing. Meaningful causal aggregation and paradoxical confounding. *arXiv preprint arXiv:2304.11625*, 2023. 10

A Background

A.1 Structural Causal Models

Causal dependencies between variables can be described using *Structural Causal Models* (SCM) (Peters et al., 2017).

Definition A.1 (SCM). *An n-dimensional structural causal model is a triplet* $\mathcal{M} = (\mathcal{G}, \mathbb{S}, P_{\boldsymbol{U}})$ *consisting of:*

- *a joint distribution* $P_{\boldsymbol{U}}$ *over exogenous variables* $\{U_j\}_{j \leq n}$,
- *a directed graph* $\mathcal{G}$ *with n vertices,*
- *a set* $\mathbb{S} = \{X_j \coloneqq f_j(\boldsymbol{Pa}_j, U_j), j = 1, \ldots, n\}$ *of structural equations, where* $\boldsymbol{Pa}_j$ *are the variables indexed by the set of parents of vertex* j *in* $\mathcal{G}$,

such that for almost every $\boldsymbol{u}$*, the system* $\{x_j \coloneqq f_j(\boldsymbol{pa}_j, u_j)\}$ *has a unique solution* $\boldsymbol{x} = \boldsymbol{g}(\boldsymbol{u})$*, with* $\boldsymbol{g}$ $P_{\boldsymbol{U}}$ *measurable.*

The unique solvability condition is included in this definition because we consider a very general class of SCMs by allowing cycles ($\mathcal{G}$ may not be a DAG). Moreover, we allow hidden confounding through the potential lack of independence between the exogenous variables $\{U_j\}$. See Bongers et al. (2016) for a thorough study of these models. Under these conditions, the distribution $P_{\boldsymbol{U}}$ entails a well-defined joint distribution over the endogenous variables $P(\boldsymbol{X})$.

Interventions in SCMs involve replacing one or more structural equations, potentially modifying exogenous distributions, and adding or removing arrows in the original graph to reflect changes in dependencies between variables. An intervention transforms the original model $\mathcal{M} = (\mathcal{G}, \mathbb{S}, P_{\boldsymbol{U}})$ into an intervened model $\mathcal{M}^{(\boldsymbol{i})} = (\mathcal{G}^{(\boldsymbol{i})}, \mathbb{S}^{(\boldsymbol{i})}, P_{\boldsymbol{U}}^{(\boldsymbol{i})})$, where $\boldsymbol{i}$ is the vector parameterizing the intervention. The base probability distribution of the unintervened model is denoted $P_{\mathcal{M}}^{(0)}(\boldsymbol{X})$ or simply $P_{\mathcal{M}}(\boldsymbol{X})$ and the interventional distribution associated with $\mathcal{M}^{(\boldsymbol{i})}$ is denoted $P_{\mathcal{M}}^{(\boldsymbol{i})}(\boldsymbol{X})$.

Classically used *do*-interventions set a structural equation to a constant, removing all influence from the parents of the intervened variable. This can be problematic for studying how the influence of low-level variables is propagated to the target, since for simultaneous interventions, the effects of some interventions can be masked by others. The probability of such masking increases as the number of low-level variables grows. *Soft interventions*, on the other hand, modify an equation while keeping the set of parents unchanged. This is more appropriate in our setting, since it propagates the information from all interventions to the target simultaneously.

Large classes of soft interventions can be designed to match domain knowledge (Besserve & Schölkopf, 2022). Notably, *shift interventions* modify the structural equation of endogenous variable l through shifting it by a scalar parameter i

$$\{X_l \coloneqq f_l(\mathbf{Pa}_l, U_l)\} \mapsto \{X_l \coloneqq f_l(\mathbf{Pa}_l, U_l) + i\}\,. \tag{4}$$

These can be combined to form multi-node interventions with vector parameter $\boldsymbol{i}$.

A.2 Causal Model Reductions (CMR)

We consider as CMR any (possibly approximate) mapping from a low-level SCM $\mathcal{L}$ to a simpler high-level SCM $\mathcal{H}$. An example is Causal Feature Learning (CFL) (Chalupka et al., 2014; 2016), which achieves a CMR by merging values of a large observation space to yield discrete high-level variables taking values in a small finite set. Consider:

- $\mathcal{L}$ has a vector of endogenous variables $\boldsymbol{X}$ with range $\mathcal{X}$ and a set of interventions $\mathcal{I}$,
- $\mathcal{H}$ has a vector of endogenous variables $\boldsymbol{Z}$ with range $\mathcal{Z}$ and a set of interventions $\mathcal{J}$.

Starting from the distribution of the low-level model $P_{\mathcal{L}}(\boldsymbol{X})$, a deterministic mapping $\tau : \mathcal{X} \to \mathcal{Z}$ generates a joint distribution on the high-level variables that is the push-forward distribution of $P_{\mathcal{L}}(\boldsymbol{X})$ by τ, denoted $\tau_{\#}[P_{\mathcal{L}}(\boldsymbol{X})]$ such that

$$\tau(\boldsymbol{X}) \sim \tau_{\#}[P_{\mathcal{L}}(\boldsymbol{X})]\,.$$

The low-level interventional distributions can be pushed forward to the high-level in the same way.

A general framework for CMR is based on the notion of *exact transformation*, which ensures *interventional consistency* by matching the push-forward low-level distributions to the high-level ones.

Definition A.2 (Exact transformation (Rubenstein et al., 2017))**.** *A map $\tau : \mathcal{X} \to \mathcal{Z}$ is an exact transformation from $\mathcal{L}$ to $\mathcal{H}$ if it is surjective, and there exists a surjective intervention map $\omega : \mathcal{I} \to \mathcal{J}$ such that for all $\boldsymbol{i} \in \mathcal{I}$*

$$\tau_{\#}[P_{\mathcal{L}}^{(i)}(\boldsymbol{X})] = P_{\mathcal{H}}^{(\omega(i))}(\boldsymbol{Z})\,.$$

The set of possible τ can be restricted to *constructive transformations*, where high-level variables depend only on non-overlapping subsets of low-level variables. This eases interpretability of CMR and comes with characterization results (Beckers & Halpern, 2019; Geiger et al., 2023a).

Definition A.3. *$\tau : \mathcal{X} \to \mathcal{Z}$ is a constructive transformation between model $\mathcal{L}$ and $\mathcal{H}$ if there exists an alignment map π relating indices of each high-level endogenous variable to a subset of indices of low-level endogenous variables such that for all $k \neq l$, $\pi(k) \cap \pi(n) = \emptyset$ and for each component τ_k of τ it exists a function $\bar{\tau}_k$ such that for all $\boldsymbol{x}$ in $\mathcal{X}$,*

$$\tau_k(\boldsymbol{x}) = \bar{\tau}_k(\boldsymbol{x}_{\pi(k)})\,.$$

The intervention map ω of constructive exact transformations are required to be constructive as well, such that acting on high-level variable k depend only on low-level interventions acting on variables in $\pi(k)$ (see App. A.5).

A.3 Numerical Schemes for Simulations

Methods for the numerical approximations of scientific models is a broad area spanning multiple fields. We provide here a few elements based on a 1D example to justify how these models relate to SCMs. The Euler method (Biswas et al., 2013), can be used to approximate a 1D ODE of the form

$$\begin{cases} x(t_0) = x_0\,, \\ \frac{dx}{dt} = F(x(t))\,, \end{cases}$$

with F smooth real-valued function, using a discretized time grid with time step Δt. The finite difference approximation of the derivative

$$x(t + \Delta t) - x(t) \approx \Delta t \cdot \frac{dx}{dt}\,,$$

leads to the iterative numerical scheme for the approximation $\hat{x}(k\Delta t)$:

$$\begin{cases} \hat{x}(t_0) = x_0\,, \\ \hat{x}(t_0 + (k+1)\Delta t) = \hat{x}(t_0 + k\Delta t) + \Delta t \cdot F(\hat{x}(t_0 + k\Delta t))\,, \quad k > 0\,. \end{cases}$$

This scheme is called *explicit* because future values depend explicitly on past ones. In that case, one may define the N low-level endogenous variables as $\boldsymbol{X} = [\hat{x}(t_0 + \Delta t), \ldots, \hat{x}(t_0 + N\Delta t)]^\top$. They can be seen as pertaining to a chain SCM with structural equations

$$\begin{cases} \widehat{X}_1 \coloneqq x_0 + \Delta t \cdot F(x_0) \\ \widehat{X}_{k+1} \coloneqq \widehat{X}_k + \Delta t \cdot F(\widehat{X}_k)\,, \quad k > 1. \end{cases}$$

Because the ODE describes deterministic dynamics, the corresponding SCM is deterministic as well, i.e. exogenous variables can be taken as trivial zero constants. However, if we turn this ODE into the following 1D SDE

$$\begin{cases} X(t_0) = x_0\,, \\ dX = F(X(t))dt + \sigma_W \cdot dW\,, \end{cases}$$

where W is a standard Brownian motion, then the Euler-Murayama method generalizes the previous approximation (Sauer, 2013), and leads to an updated SCM with structural equation

$$\begin{cases} \widehat{X}_1 \coloneqq x_0 + \Delta t \cdot F(x_0) + U_1 \\ \widehat{X}_{k+1} \coloneqq \widehat{X}_k + \Delta t \cdot F(\widehat{X}_k) + U_{k+1}\,, \quad k > 1\,, \end{cases}$$

where the exogenous variables U_k represent the increments of the scaled Brownian motion $\sigma_W \cdot W$ between successive time steps, and are thus jointly independent Gaussian due to fundamental properties of Brownian motion.

This approach generalizes to explicit numerical schemes for multivariate ODE and SDEs where the state variable $\boldsymbol{X}$ lives in $\mathbb{R}^n$. As an illustration, we can take the following class of SDE models

$$\begin{cases} \boldsymbol{X}(t_0) = \boldsymbol{x}_0\,, \\ d\boldsymbol{X} = \boldsymbol{F}(\boldsymbol{X}(t))dt + \boldsymbol{\sigma}_W \cdot d\boldsymbol{W}\,, \end{cases}$$

with $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^n$, $\boldsymbol{\sigma}_W = \mathbb{R}^{(n\times n)}$, and $\boldsymbol{W}$ a n-dimensional standard Brownian motion. This leads to the scheme

$$\begin{cases} \widehat{\boldsymbol{X}}_1 \coloneqq \boldsymbol{x}_0 + \Delta t \cdot \boldsymbol{F}(\boldsymbol{x}_0) + \boldsymbol{U}_1 \\ \widehat{\boldsymbol{X}}_{k+1} \coloneqq \widehat{\boldsymbol{X}}_k + \Delta t \cdot \boldsymbol{F}(\widehat{X}_k) + \boldsymbol{U}_{k+1}\,, \quad k > 1\,, \end{cases}$$

where the $\boldsymbol{U}_k$ are now multivariate Gaussian variables, whose components may or may not be independent depending on the choice of the matrix $\boldsymbol{\sigma}_W$. If the exogenous components are independent, the variables can be described by a standard SCM as introduced in the main text. If the exogenous components are dependent, the variables can be described by a more general notion of SCM, allowing hidden confounding (Bongers et al., 2021).

Further generalization to numerical schemes for Stochastic Partial Differential Equations (SPDEs) using finite difference approximations for partial derivatives with respect to other variables than time are also possible (Millet & Morien, 2005).

A.4 Reduction of the Euler Scheme for a System of Point Masses

In the context of the main text example, we assume each point Mass is submitted to a fluid friction force opposing its movement with fixed coefficient λ. Masses are moreover intervened on via additional external forces $\{\boldsymbol{f}_k\}$. Finally, internal forces are exerted on mass k by other point masses of the system, summing up to $\boldsymbol{g}_k$. Newton's second law applied to individual masses results in the following system of 2D vector equations

$$m_k \frac{d\boldsymbol{v}_k}{dt} = -\lambda \boldsymbol{v}_k(t) + \boldsymbol{f}_k(t) + \boldsymbol{g}_k(t)\,.$$

We can approximate each equation to estimate interatively the x and y components of the speed of individual point masses in the system, using a small time-step Δt, such that we get the discrete time estimates $\hat{v}_{x,k}[n] \approx v_{x,k}(n\Delta t)$ and $\hat{v}_{y,k}[n] \approx v_{y,k}(n\Delta t)$ satisfying

$$\begin{aligned} m_k \cdot \hat{v}_{x,k}[n+1] &\coloneqq (1 - \Delta t\lambda) \cdot m_k \cdot \hat{v}_{x,k}[n] + \Delta t \cdot f_{x,k}(n\Delta t) + \Delta t \cdot g_{x,k}(n\Delta t)\,, \\ m_k \cdot \hat{v}_{y,k}[n+1] &\coloneqq (1 - \Delta t\lambda) \cdot m_k \cdot \hat{v}_{y,k}[n] + \Delta t \cdot f_{y,k}(n\Delta t) + \Delta t \cdot g_{y,k}(n\Delta t). \end{aligned}$$

Here, f represents external forces, λ is a viscous damping coefficient, and g denotes internal forces. We consider $\boldsymbol{i}$ to be the vector of all components of external forces, and the target variable to be the final horizontal speed of the center of mass at iteration N. From the physics of freely moving systems of points, it is clear that the target variable can be predicted by considering only the horizontal dynamics of the center of mass. More precisely, we integrate the sum of external forces over the time span of the experiment, and use the last intervened time point n_f to predict the final outcome of the simulation, leading to the reduction

$$Z_1^{(\omega(\boldsymbol{i}))} = \left(\sum_k m_k\right) v_{x,G}[n_f] = \sum_k m_k v_{x,k}[n_f] = \Delta t \sum_{n=0}^{n_f} (1 - \Delta t\lambda)^{(n_f - n)} \sum_k f_{x,k}(n\Delta t))\,,$$

$$Y = v_{x,G}[N] \coloneqq (1 - \Delta t\lambda)^{(N - n_f)}\, Z_1 + \sum_{n=n_f+1}^{N} \sum_k f_{x,k}(n\Delta t))\,.$$

To make the notation compatible with that used in our TCR framework, we can gather all speed variables in a high-dimensional vector $\boldsymbol{X}$ and all external force variables in a vector $\boldsymbol{i}$, the high-level causal model is thus generated by a linear τ-map and linear ω map for shift interventions, taking the form of the exact transformation

$$Z_1 = \boldsymbol{\tau}_1^\top \boldsymbol{X} + \boldsymbol{\omega}_1^\top \boldsymbol{i}\,, \quad Y = \boldsymbol{\tau}_0^\top \boldsymbol{X} + \boldsymbol{\omega}_0^\top \boldsymbol{i} \coloneqq f(Z_1)\,.$$

where the term $\boldsymbol{\omega}_0^\top \boldsymbol{i}$ accounts for interventions happening between discrete times $n_f + 1$ and N and thus affect Y without being mediated by Z_1. In our framework, only interventions mediated by the cause, reflected in the term $\boldsymbol{\omega}_1^\top \boldsymbol{i}$, are accounted for in the high-level model.

A.5 Constructive Transformations

We complete the main text definition to include the constraint on the intervention map ω

Definition A.4. $(\tau, \omega) : (\mathcal{X} \to \mathcal{Z}, \mathcal{I} \to \mathcal{J})$ *is a constructive* $(\tau - \omega)$*-transformation between model* $\mathcal{L}$ *and* $\mathcal{H}$ *if there exists an alignment map* π *mapping each high-level endogenous variable to a subset of low-level endogenous variables such that for all* $k \neq l$, $\pi(k) \cap \pi(n) = \emptyset$ *and we have both*

- *for each component* τ_k *of* τ *there exists a function* $\bar{\tau}_k$ *such that for all* $\boldsymbol{x}$ *in* $\mathcal{X}$,

$$\tau_k(\boldsymbol{x}) = \bar{\tau}_k(\boldsymbol{x}_{\pi(k)}) \, ;$$

- *for each component* ω_k *of* ω *there exists a function* $\bar{\omega}_k$ *such that for all* $\boldsymbol{i}$ *in* $\mathcal{I}$,

$$\omega_k(\boldsymbol{i}) = \bar{\omega}_k(\boldsymbol{i}_{\pi(k)}) \, .$$

A.6 Related Work

Desiderata for CMR have been addressed theoretically by several works, in particular in the context of CFL (Chalupka et al., 2014; 2016), and subsequently with the notion of *exact transformations* (Rubenstein et al., 2017) and a more strongly constrained subclass: *causal abstractions* (Beckers & Halpern, 2019). An alternative framework for composing abstractions of finite models has been proposed by Rischel & Weichwald (2021). However, only few works have addressed how to build high-level representations from the low-level system data only. A line of works focuses on language models (Geiger & Straehle, 2020; Geiger et al., 2021; 2023a), where high-level variables and interpretations are readily available.

We start from the opposite direction and develop a general approach to build the high-level abstraction from the ground-up. Such a construction is done in CFL (Chalupka et al., 2014; 2016), where high-dimensional microscopic variables are turned into discrete high-level variables. Zennaro et al. (2023) addressed this question in the context of finite and discrete domains, by minimizing the maximum Jensen-Shannon divergence over a finite set of perfect intervention distributions. In contrast with most works on CMR, our framework is fully compatible with imperfect (soft intervention) at the low level, which are more realistic and interpretable perturbations of many real-world systems than hard interventions. Soft interventions have been used for language model alignment (Geiger et al., 2023b), and their theoretical compatibility with the abstraction framework has been investigated by Massidda et al. (2023). Our approach aims at approximating an exact transformation, and is thus a relaxation of this setting.

Other theoretical frameworks for approximate abstractions have been proposed (Beckers et al., 2020; Rischel & Weichwald, 2021). Our work differs by providing an explicit loss well-suited to continuous causal models, that can be optimized efficiently and provide interpretable outcomes thanks to a cause-mechanism decomposition, a lower bound, and analytic solutions.

The way we relax constraints on low-level interventions shares also similarities with the views of Zhu et al. (2023) who consider stochastic low-level do-interventions sampled according to the observational distribution, our work is instead focused on soft-interventions, for which we impose a prior distribution, reflecting the relative importance that we put on them. Our optimization objective is averaged over this prior, such that it plays a role in the final solution.

Our approach also relates to the search for optimal interventional or counterfactual manipulations to steer the output of a system to a particular value or distribution (Amos et al., 2018; Besserve et al., 2020) or to best explain an observation (Budhathoki et al., 2022; Von Kügelgen et al., 2023). We are in a way also selecting particular manipulations, but through the choice of dimensionality reduction ω, such that they are interpretable at a high level.

Finally, our approach relates to several works in causal representation learning, which have addressed identifiability of latent causal models from observational data, with (Liang et al., 2023) or without (Squires et al., 2023; von Kügelgen et al., 2023) assumptions on the latent causal graph. In contrast

to those works, TCR does not assume an injective mapping of the mapping of the observations to the latent variables, such that the high-level model typically losses information relative to the low-level model.

B Theoretical Analysis

As described in Fig. 1b, we consider endogenous variables of a low-level model gathered in a (high-dimensional) random vector $\boldsymbol{X}$. A target scalar variable $Y = \tau_0(\boldsymbol{X})$ quantifies a property of interest of this model, and can be thought of as quantifying the presence or magnitude of a *phenomenon* in the data, using *detector* τ_0. To generate a high-level causal explanation of this phenomenon, we learn a high-level SCM with a fixed causal structure, where the known effect variable Y is caused by n learned independent high-level variables Z_k. The low-level variables $\boldsymbol{X}$ are approximately mapped to the high-level variables $\boldsymbol{Z}$ using a constructive transformation and an associated constructive interventional map with the same alignment π.

B.1 TCR Framework

Our reduction framework has the following elements:

(1) A low-level SCM $\mathcal{L}$ with N endogenous variables $\{X_1, X_2, \dots, X_N\}$ and corresponding exogenous variables $\{U_k\}_{k=1..N}$ equipped with joint distribution $P(\boldsymbol{U})$. A set of low-level interventions parameterized by vector $\boldsymbol{i} \in \mathcal{I}$ with distribution $P(\boldsymbol{i})$, with each component i_k affecting a unique endogenous variable X_k. We only assume we can query (samples form) unintervened and interventional distributions of $\mathcal{L}$.

(2) A class of high-level SCMs $\{\mathcal{H}_{\boldsymbol{\gamma}}\}_{\gamma\in\Gamma}$ with (n+1) endogenous variables $\{Y, Z_1, \dots, Z_n\}$ and associated exogenous variables $\{R_k\}_{k=0..n}$, equipped with a factorized distribution $P(\boldsymbol{R}) = \prod P_{R_k}$. A set of high-level interventions parametrized by vector $\boldsymbol{j} \in \mathcal{J}$, with each component j_k affecting a single node Z_k. In contrast to the (fixed) low-level model, the high-level model parameters $\boldsymbol{\gamma}$ need to be learned.

These two models are linked by a constructive transformation with two deterministic surjective maps τ and ω from low- to high-level endogenous variables and interventions, respectively, which decompose as

$$\tau = (\tau_0, \tau_1, \tau_2, \dots, \tau_n) \text{ with } \tau_k : x \mapsto \bar{\tau}_k(x_{\pi(k)}) \tag{5}$$

$$\omega = (\omega_0, \omega_1, \omega_2, \dots, \omega_n) \text{ with } \omega_k : \boldsymbol{i} \mapsto \bar{\omega}_k(\boldsymbol{i}_{\pi(k)}) \tag{6}$$

where π is a so-called alignment function from $[0\,..\,n]$ to non-overlapping subsets of $[1\,..\,N]$. Importantly, τ_0 (and thus $(\bar{\tau}_0, \pi(0))$) are assumed fixed and known. Additionally, ω_0 is assumed to be a trivial constant map $\boldsymbol{i} \to 0$, to ensure that the high-level target variable cannot be directly intervened upon, as we want to explain the changes in Y exclusively through changes of its high-level causes.

The high-level model involves the following mechanisms which need to be learned: (1) The marginal distribution of each high-level cause $P^{(\boldsymbol{j})}(Z_k)$ in all high-level interventional settings $\boldsymbol{j}$. (2) The mechanism $P(Y|\boldsymbol{Z})$ mapping high-level causes to Y, comprised of the distribution of the exogenous variable R_0 and the map

$$(Z_1, ..., Z_n, R_0) \mapsto f_{\boldsymbol{\gamma}}(Z_1, ..., Z_n, R_0) =: Y\,.$$

B.2 Causal Consistency Loss

It is not always possible to achieve an exact transformation that guarantees consistency of low- and high-level models for almost all interventions. As a consequence, we allow for the consistency between models to be approximate. To ensure that this approximation is as accurate as possible, we minimize the expected KL divergence between the pushforward by the transformation τ of the low-level interventional distributions that we denote $\widehat{P}_{\tau}^{(\boldsymbol{i})}(Y, \boldsymbol{Z}) = \tau_{\#}[P_{\mathcal{L}}^{(\boldsymbol{i})}(\boldsymbol{X})]$, and the corresponding interventional distribution of the high-level model $P^{(\omega(\boldsymbol{i}))}$, leading to the consistency loss

$$\mathcal{L}_{cons} = \mathbb{E}_{\boldsymbol{i}\sim P(\boldsymbol{i})}\left[KL\left(\widehat{P}_{\tau}^{(\boldsymbol{i})}(Y, \boldsymbol{Z}) \| P^{(\omega(\boldsymbol{i}))}(Y, \boldsymbol{Z})\right)\right]\,. \tag{7}$$

Other losses have been previously suggested to enforce consistency. Beckers et al. (2020) propose to take a maximum over interventions, whereas we take the expectation in our loss, thus focussing the CMR on the average performance rather than the worst case. Rischel & Weichwald (2021); Zennaro et al. (2023) use the Jensen-Shannon (JS) divergence in the context of finite models. Instead, we choose the KL divergence because, contrary to JS, it leads to a tractable expression under Gaussian assumptions. Moreover, the proposed consistency loss (7) has the following properties.

Proposition B.1 (Consistency loss)**.** *The consistency loss is positive, invariant to invertible reparametrizations (see Def. E.1), and vanishes if and only if the transformation is exact for almost all interventions. It decomposes as*

$$\mathcal{L}_{cons} = \mathbb{E}_{i\sim P(\boldsymbol{i})}\Bigg[KL\left(\widehat{P}_{\tau}^{(\boldsymbol{i})}\left(\boldsymbol{Z}\right)||P^{(\omega(\boldsymbol{i}))}\left(\boldsymbol{Z}\right)\right) + \mathbb{E}_{z\sim\widehat{P}_{\tau}^{(\boldsymbol{i})}(\boldsymbol{Z})}\Big[KL\left(\widehat{P}_{\tau}^{(\boldsymbol{i})}\left(Y|z\right)\right)||P^{(0)}\left(Y|z\right)\right)\Big]\Bigg], \quad (8)$$

and is an upper bound of the causal relevance loss

$$\mathcal{L}_{rel} = \mathbb{E}_{\boldsymbol{i}\sim P(\boldsymbol{i})}\left[KL\left(\widehat{P}^{(\boldsymbol{i})}\left(Y\right)||P^{(\omega(\boldsymbol{i}))}\left(Y\right)\right)\right] \leq \mathcal{L}_{cons}\,. \quad (9)$$

Reparametrization invariance (see Def. E.1) refers to transformations of the pairs $(\tau, f_{\boldsymbol{\gamma}})$ that leave the composition $f_{\boldsymbol{\gamma}} \circ \tau$ invariant. In the $n = 1$ linear setting (see Sec. B.3), this corresponds to invariance by multiplicative rescaling. This guarantees that equivalent high-level causal descriptions treated equally by the loss.

We call Eq. (8) a *Cause-Mechanism Decomposition* because the first term quantifies the *cause consistency* and the second term can be thought of as the *mechanism consistency*. This latter term assesses the similarity between the outputs of the learned high-level mechanism $P^{(0)}(Y|z)$ and the corresponding conditional distribution computed by push-forward of the low-level variables $\widehat{P}_{\tau}^{(\boldsymbol{i})}(Y|z)$. Since we prevent the high-level mechanism from being intervened on, only its unintervened conditional appears in the expression.

Lastly, the causal relevance loss $\mathcal{L}_{rel}$ assesses whether the variations of the target Y due to low-level interventions are well-captured by high-level interventions, on average over the prior $P(\boldsymbol{i})$. Its upper bound by $\mathcal{L}_{cons}$ ensures that by optimizing for consistency, we also indirectly promote effective "explanation" of the variations in the target density resulting from low-level interventions. We can thus choose $P(\boldsymbol{i})$ to make the most relevant interventions more likely according to domain knowledge, such that optimizing the loss will steer towards a solution capturing the most domain-relevant variations of the target.

B.3 Linear Reduction with Shift Interventions

We further constrain the setting to be able to study the solution minimizing $\mathcal{L}_{cons}$ analytically and get insights into the properties of TCR.

Notations. We use boldface for column vectors, and $\boldsymbol{i}_S$ for the subvector of $\boldsymbol{i}$ restricted to the components in set S. When a vector, say $\boldsymbol{\tau}_k$, is associated to a high-level SCM component k of a constructive transformation with alignment π, $\bar{\boldsymbol{\tau}}_k$ indicates the restriction of $\boldsymbol{\tau}_k$ to components in $\pi(k)$. The number of elements in a set S is denoted $\#S$.

Tau map. To maximize interpretability, we assume a linear τ-map, represented as a vector $\boldsymbol{\tau}$ such that:

$$\boldsymbol{X} \mapsto \begin{bmatrix} Y \\ \boldsymbol{Z} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\tau}_0^\top \\ \vdots, \\ \boldsymbol{\tau}_n^\top \end{bmatrix} \boldsymbol{X} = \left[\bar{\boldsymbol{\tau}}_0^\top \boldsymbol{X}_{\pi(0)}\,, \quad \ldots \bar{\boldsymbol{\tau}}_n^\top \boldsymbol{X}_{\pi(n)}\right]^\top.$$

Omega map. We focus on *shift interventions* and map the vector $\boldsymbol{i}$ of low-level interventions on the nodes in $\pi(k)$ to a scalar shift intervention on the mechanism of each Z_k. We assume each map ω_k

to be linear with vector $\boldsymbol{\omega}_k$ such that

$$\omega_k(\boldsymbol{i}) = \boldsymbol{\omega}_k^\top \boldsymbol{i} = \bar{\boldsymbol{\omega}}_k^\top \boldsymbol{i}_{\pi(k)} \,.$$

Because high-level causes are root nodes, intervening amounts to shifting the marginal distribution from $P^{(0)}(Z_k)$ to $P^{(\omega_k(\boldsymbol{i}))}(Z_k) = P^{(0)}(Z_k - \omega_k(\boldsymbol{i}))$.

Choice of alignment π. There are potential degrees of freedom for π, and users may want to incorporate domain knowledge as well as interpretability constraints to reduce the variables included in $\cup_{k\neq 0}\pi(k)$. In practice, we learn the distribution of the low-level variables among the $\pi(k)$ using regularization (see Sec. C).

Choice of prior $P(\boldsymbol{i})$. The solutions minimizing the loss of Eq. (7), may depend on the choice of the prior $P(\boldsymbol{i})$, and in particular on which variables are actually intervened on. Let Ω denote the subset of indices of low-level variables that are intervened on with non-zero probability. The components of $\boldsymbol{i}$ whose index does not belong to Ω thus take value $i = 0$ with probability one. We provide identifiability guaranties under two kinds of assumptions.

Assumption B.2. $P(\boldsymbol{i}_\Omega)$ has a density with respect to the Lebesgue measure, with support covering a neighborhood of zero (*i.e.* the unintervened case).

Assumption B.3. There are at least #Ω distinct interventions happening with non-zero probability. Corresponding to a family of values of the vector of $\boldsymbol{i}_\Omega$, with full rank #Ω.

While Assum B.2 gives a simple way to draw interventions from intuitive prior densities that reflect the knowledge on how likely those are, Assum B.3 allows addresses a classical question in causal representation learning: *How many distinct interventions are needed to learn the reduction?*

B.4 Identifiability Results

If we assume the low-level model is linear Gaussian of the form $\boldsymbol{X}_\Omega \to \boldsymbol{X}_{\pi(0)}$, we can show the existence and uniqueness of the solution.

Proposition B.4. *Assume the low-level SCM follows*

$$\boldsymbol{X} \coloneqq A\boldsymbol{X} + \boldsymbol{U} + \boldsymbol{i}\,, \quad U_k \sim \mathcal{N}(\mu_k, \sigma_k^2)\,,\ \boldsymbol{i} \sim P(\boldsymbol{i})$$

such that $\boldsymbol{X}$ and A take the block forms

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_{\pi(0)} \\ \boldsymbol{X}_\Omega \end{bmatrix}, \quad A = \begin{bmatrix} A_{00} & A_{0\Omega} \\ \boldsymbol{0} & A_{\Omega\Omega} \end{bmatrix}.$$

Given an arbitrary choice of linear scalar target of the form $Y = \tau_0^\top \boldsymbol{X} = \bar{\tau}_0^\top \boldsymbol{X}_{\pi(0)}$ and under Assum. B.2 or Assum. B.3, there is a unique linear 1-cause TCR (up to a multiplicative constant) satisfying $\mathcal{L}_{cons} = 0$. It is given by

$$\bar{\boldsymbol{\tau}}_1 = A_{01}^\top (I_{\#\pi(0)} - A_{00})^{-\top} \bar{\boldsymbol{\tau}}_0 \tag{10}$$

$$\text{and} \quad \bar{\boldsymbol{\omega}}_1 = (I_{\#\Omega} - A_{\Omega\Omega})^{-\top} \bar{\boldsymbol{\tau}}_1 \,. \tag{11}$$

Moreover, let n_{max} be the maximum number such that a linear n-cause TCR can achieve $\mathcal{L}_{cons} = 0$. If there is no cancellations among causal pathways from each node in supp$(\bar{\boldsymbol{\omega}}_1)$ of Eq. (11) towards target Y, then the n_{max}-cause TCR is unique up to rescaling and permutation of the causes.

This result provides guaranties for having a unique ground-truth solution in case exact transformations can be achieved. The main assumption is the absence of feedback influences from the target set $\pi(0)$ to candidate causes. However, cycles and confounding are allowed in the low-level model, contrary to the learned high-level model. The 1-cause solution is easiest to obtain. The study of simple SCMs (App. E.2 and App. E.3) provides some insights on the form of the analytical solution. Additional results show that we lose identifiability of the TCR if we drop the assumption that not all variables in $\pi(1)$ are intervened on (see G.1). The n-cause solution is essentially a partition of the 1-cause solution that enforce independence between them. By cancellation of causal pathways, we mean that there exists a subset of oriented paths from A to B such that their total effect cancels out. Note that it does not prevent A to have an total effect on B. Assuming non-cancellation of causal paths is akin to preventing faithfulness violations and generically satisfied.

Algorithm 1 Linear TCR (LCPR)

Input λ: learning rate, $P(\boldsymbol{i})$: intervention prior, $Simulate(\boldsymbol{\theta}, \boldsymbol{i}, n_{\text{sim}})$: function returning n_{sim} paths, N_{ite}: number of iterations, B: simulation paths batch size, $B_{\boldsymbol{i}}$: intervention batch size.
Output Estimated parameters $(\boldsymbol{\tau}_1, \boldsymbol{\omega}_1, \boldsymbol{\gamma})$.
Initialize $\boldsymbol{\tau}_1, \boldsymbol{\omega}, \boldsymbol{\gamma}$

```
for m = 1..N_ite do
    X, Y ← []
    for l = 1..B_i do
        i_l ← Sample(P(i))
        X_l = (x^1, .., x^B), Y_l ← Simulate(θ, i_l, B)
        X ← [X[:], X_l]
        Y ← [Y[:], Y_l]
        I = [I[:], i_l]
    L_tot ← ComputeLoss(X, Y, I, τ_1, ω_1, γ)
    ∇_γ, ∇_τ ← ComputeLossGradient(L_tot)
    (γ, τ_1, ω_1) ← (γ − λ∇_γ, τ_1 − λ∇_{τ_1}, ω_1 − λ∇_{ω_1})
```

C Linear TCR Algorithm

In this section, we introduce an algorithm to learn a linear reduction with shift interventions.

Gaussian approximation of consistency loss. Since the KL divergence is challenging to learn non-parametrically, we make a Gaussian assumption on the densities. This allows us to obtain an analytic expression for the loss based on second order statistics (see expression in App. F.1).

Overlap loss. To ensure differentiability of the reduction maps we do not implement the alignment π explicitly, but encourage non-overlapping reduction maps via the regularizer

$$\mathcal{L}_{\text{ov}} = \sum_{k<l} \left(\left\langle \frac{|\boldsymbol{\tau}_k|}{\|\boldsymbol{\tau}_k\|}, \frac{|\boldsymbol{\tau}_l|}{\|\boldsymbol{\tau}_l\|} \right\rangle + \left\langle \frac{|\boldsymbol{\omega}_k|}{\|\boldsymbol{\omega}_k\|}, \frac{|\boldsymbol{\omega}_l|}{\|\boldsymbol{\omega}_l\|} \right\rangle \right), \tag{12}$$

where $|\cdot|$ is the element-wise absolute value.

Balancing loss. Minimizing the Gaussian approximation of the consistency loss together with overlap regularization (12) there is nothing preventing the solution from attributing all non-zero weights in the τ- and ω maps to one high-level variable while ignoring all others. In order to prevent such a collapse, we minimize stark differences between the high-level variables through the balancing term

$$\mathcal{L}_{\text{bal}} = \left(\frac{\sqrt{\sum_k \|\alpha_k \boldsymbol{\tau}_k\|^2}}{\sum_k \|\alpha_k \boldsymbol{\tau}_k\|^2} + \frac{\sqrt{\sum_k \|\alpha_k \boldsymbol{\omega}_k\|^2}}{\sum_k \|\alpha_k \boldsymbol{\omega}_k\|^2} \right), \tag{13}$$

where α_k is the coefficient in the linear high-level mechanism corresponding to variable Z_k.

Gathering the losses, we get the total objective

$$\underset{\boldsymbol{\gamma}, \boldsymbol{\tau}, \boldsymbol{\omega}}{\text{minimize}} \; \mathcal{L}_{\text{tot}} = \mathcal{L}_{\text{cons}} + \eta_{\text{ov}} \mathcal{L}_{\text{ov}} + \eta_{\text{bal}} \mathcal{L}_{\text{bal}} \,. \tag{14}$$

The learning procedure is described in Algorithm 1.

D Proofs

D.1 Proof of Proposition B.1

We first reformulate Proposition B.1 more formally as follows.

Proposition D.1. *The consistency loss is positive, invariant to invertible reparametrizations as defined in Definition E.1, and vanishes if and only if the transformation is exact for almost all*

interventions. It admits the following decomposition:

$$\mathcal{L}_{cons} = \mathbb{E}_{i\sim P(\boldsymbol{i})}\left[KL\left(\widehat{P}^{(\boldsymbol{i})}_{\tau}\left(\boldsymbol{Z}\right)||P^{(\omega(\boldsymbol{i}))}\left(\boldsymbol{Z}\right)\right) + \mathbb{E}_{z\sim\widehat{P}^{(\boldsymbol{i})}_{\tau}(\boldsymbol{Z})}\left[KL\left(\widehat{P}^{(\boldsymbol{i})}_{\tau}\left(Y|z\right)\right)||P^{(0)}\left(Y|z\right)\right)\right]\right], \quad (15)$$

and is an upper bound of the causal relevance loss

$$\mathcal{L}_{rel} = \mathbb{E}_{\boldsymbol{i}\sim P(\boldsymbol{i})}\left[KL\left(\widehat{P}^{(\boldsymbol{i})}\left(Y\right)||P^{(\omega(\boldsymbol{i}))}\left(Y\right)\right)\right] \leq \mathcal{L}_{cons}\,. \quad (16)$$

Proof. **Positivity** of the loss comes from the positivity of the KL-divergence. Taking the expectation of this divergence with respect to $P(\boldsymbol{i})$ thus must be positive too.

Invariance to reparameterizations. We assume a reparametrization ρ designed according to the framework introduced in Appendix E.1. By invariance of the KL divergence to invertible transformations, we have equality between the KL associated to the two different reductions (τ, ω) and $(\rho\circ\tau, \psi\circ\omega)$:

$$KL\left(\widehat{P}^{(\boldsymbol{i})}_{\tau}(Y,\boldsymbol{Z})||P^{(\omega(\boldsymbol{i}))}_{\mathcal{H},\boldsymbol{\gamma}}(Y,\boldsymbol{Z})\right) = KL\left(\tilde{\rho}_{\#}[\widehat{P}^{(\boldsymbol{i})}_{\tau}(Y,\boldsymbol{Z})]||\tilde{\rho}_{\#}[P^{(\omega(\boldsymbol{i}))}_{\mathcal{H},\boldsymbol{\gamma}}(Y,\boldsymbol{Z})]\right) = KL\left(\widehat{P}^{(\boldsymbol{i})}_{\rho\circ\tau}(Y,\boldsymbol{Z})||P^{(\psi\circ\omega(\boldsymbol{i}))}_{\mathcal{H},\boldsymbol{\gamma}'}(Y,\boldsymbol{Z})\right).$$

The transformation (ρ, ψ) thus leaves $\mathcal{L}_{cons}$ invariant.

Cause-mechanism decomposition. Under our setting (see Sec. B.1), the interventional distribution of the high-level causal model factorizes as

$$P^{(\omega(\boldsymbol{i}))}(Y,\boldsymbol{Z}) = P^{(0)}\left(Y|\boldsymbol{Z}\right)P^{(\omega(\boldsymbol{i}))}\left(\boldsymbol{Z}\right).$$

The pushforward (by reduction) of the interventional distribution of the low-level model factorizes as

$$\widehat{P}^{(\boldsymbol{i})}(Y,\boldsymbol{Z}) = \widehat{P}^{(\boldsymbol{i})}\left(Y|\boldsymbol{Z}\right)\widehat{P}^{(\boldsymbol{i})}\left(\boldsymbol{Z}\right),$$

with $\widehat{P}^{(\boldsymbol{i})}\left(\boldsymbol{Z}\right) = \tau_{1,\#}[P^{(\boldsymbol{i})}(\boldsymbol{X}_{\pi(1)})]$ and $\widehat{P}^{(\boldsymbol{i})}\left(Y|\boldsymbol{Z}\right) = \frac{\tau_{\#}\left[P^{(\boldsymbol{i})}\left(\boldsymbol{X}_{\pi(0)},\boldsymbol{X}_{\pi(1)}\right)\right]}{\tau_{1,\#}[P^{(\boldsymbol{i})}(\boldsymbol{X}_{\pi(1)})]}$.

Thus, the KL divergence can be decomposed as

$$\begin{aligned}
&KL\left(\widehat{P}^{(\boldsymbol{i})}(Y,\boldsymbol{Z})||P^{(\omega(\boldsymbol{i}))}(Y,\boldsymbol{Z})\right)\\
&= \int_{\mathcal{Y}}\int_{\mathcal{Z}}\widehat{P}^{(\boldsymbol{i})}(Y,\boldsymbol{Z})\log\frac{\widehat{P}^{(\boldsymbol{i})}(Y,\boldsymbol{Z})}{P^{(\omega(\boldsymbol{i}))}(Y,\boldsymbol{Z})}d\boldsymbol{Z}dY\\
&= \int_{\mathcal{Y}}\int_{\mathcal{Z}}\widehat{P}^{(\boldsymbol{i})}\left(Y|\boldsymbol{Z}\right)\widehat{P}^{(\boldsymbol{i})}\left(\boldsymbol{Z}\right)\log\frac{\widehat{P}^{(\boldsymbol{i})}\left(Y|\boldsymbol{Z}\right)\widehat{P}^{(\boldsymbol{i})}\left(\boldsymbol{Z}\right)}{P^{(0)}\left(Y|\boldsymbol{Z}\right)P^{(\omega(\boldsymbol{i}))}\left(\boldsymbol{Z}\right)}d\boldsymbol{Z}dY\\
&= KL_{Z}\left(\widehat{P}^{(\boldsymbol{i})}\left(\boldsymbol{Z}\right)||P^{(\omega(\boldsymbol{i}))}\left(\boldsymbol{Z}\right)\right) + \mathbb{E}_{z\sim\widehat{P}^{(\boldsymbol{i})}(\boldsymbol{Z})}\left[KL_{Y}\left(\widehat{P}^{(\boldsymbol{i})}\left(Y|\boldsymbol{Z}=z\right)||P^{(0)}\left(Y|\boldsymbol{Z}=z\right)\right)\right]\\
&= KL_{Z}\left(\widehat{P}^{(\boldsymbol{i})}\left(\boldsymbol{Z}\right)||P^{(\omega(\boldsymbol{i}))}\left(\boldsymbol{Z}\right)\right) + KL_{Y,Z}\left(\widehat{P}^{(\boldsymbol{i})}\left(\widehat{Y},\boldsymbol{Z}\right)||P^{(0)}\left(Y|\boldsymbol{Z}\right)\widehat{P}^{(\boldsymbol{i})}\left(\boldsymbol{Z}\right)\right).
\end{aligned}$$

The first term is a cause consistency loss (same principle but for that variable only). The second term can be thought of as a mechanism consistency, where we use the ground truth low-level cause distribution to probe the similarity of the outputs of the "true" (in fact, the conditional distribution) and approximate mechanism. Our interpretability choice prevents the high-level mechanism from being intervened on, so a single stochastic map (i.e. a Markov kernel) must fit at best all the sampled experimental conditionals.

Lower bounding by causal relevance We may ask the question of causal relevance of high-level causes. One way to quantify this is to assess whether the variations of the target due to low-level interventions are well captured by high-level interventions, which can be measured by a KL divergence on the target's marginal

$$\mathcal{L}_{rel} = \mathbb{E}_{\boldsymbol{i}\sim p(\boldsymbol{i})}\left[KL_{Y}\left(\widehat{P}^{(\boldsymbol{i})}\left(Y\right)||P^{(\omega(\boldsymbol{i}))}\left(Y\right)\right)\right].$$

Note: In the Gaussian 1D case, this gives

$$\frac{1}{2}\mathbb{E}_{\boldsymbol{i}\sim p(\boldsymbol{i})}\left[\frac{(\mu_Y+\alpha\boldsymbol{\omega}^\top\boldsymbol{i}-\widehat{\mu}_Y^{(\boldsymbol{i})})^2}{\sigma_Y^2}+\frac{\widehat{\sigma^2}_Y^{(\boldsymbol{i})}}{\sigma_Y^2}-\ln\left(\frac{\widehat{\sigma^2}_Y^{(\boldsymbol{i})}}{\sigma_Y^2}\right)-1\right].$$

Interestingly, we can break down this term using

$$\begin{aligned}&KL\left(\widehat{P}^{(\boldsymbol{i})}(Y,\boldsymbol{Z})||P^{(\omega(\boldsymbol{i}))}(Y,\boldsymbol{Z})\right)\\&\quad=KL_Y\left(\widehat{P}^{(\boldsymbol{i})}(Y)\,||P^{(\omega(\boldsymbol{i}))}(Y)\right)+\mathbb{E}_{y\sim\widehat{P}^{(\boldsymbol{i})}(Y)}\left[KL_Y\left(\widehat{P}^{(\boldsymbol{i})}(\boldsymbol{Z}|Y=y)\,||P^{(\omega(\boldsymbol{i}))}(\boldsymbol{Z}|Y=y)\right)\right]\end{aligned}$$

where both terms are positive by positivity of the KL divergence. As a consequence,

$$\begin{aligned}KL_Y\left(\widehat{P}^{(\boldsymbol{i})}(Y)\,||P^{(\omega(\boldsymbol{i}))}(Y)\right)&=KL\left(\widehat{P}^{(\boldsymbol{i})}(Y,\boldsymbol{Z})||P^{(\omega(\boldsymbol{i}))}(Y,\boldsymbol{Z})\right)-\mathbb{E}_{y\sim\widehat{P}^{(\boldsymbol{i})}(Y)}\left[KL_Y\left(\widehat{P}^{(\boldsymbol{i})}(\boldsymbol{Z}|Y=y)\,||P^{(\omega(\boldsymbol{i}))}(\boldsymbol{Z}|Y=y)\right)\right]\\&\leq KL\left(\widehat{P}^{(\boldsymbol{i})}(\widehat{Y},\boldsymbol{Z})||P^{(\omega(\boldsymbol{i}))}(Y,\boldsymbol{Z})\right)=\mathcal{L}_{cons}\,.\end{aligned}$$

so the minimized consistency loss is an upper bound to causal relevance.

□

D.2 Proof of Proposition B.4

In addition to the proof of this proposition, we are going to show that Assum. B.2 can be replaced by the following Assum. B.3 in this proposition, to yield the same result.

Proof. We exploit the positive definiteness of the KL loss and its continuity with respect to $\boldsymbol{i}$. Since the variables are jointly Gaussian, continuity is obvious from the analytical expression of the KL for Gaussian variables and continuity of the shift operation applied to the parameters of the Gaussian. We exploit the cause-mechanism decomposition and the lower-bound by $\mathcal{L}_{cons}$ to progressively identify necessary conditions on parameters to have $\mathcal{L}_{cons}=0$ and finally check those conditions are sufficient.

Let N_0 be the size of $\pi(0)$ and N_1 be the size of $\pi(1)$. Because of the SCM assumption, the causal graph is a DAG, such that variables can be ordered without loss of generality such that A (and thus A_{00} and A_{11}) is strictly upper triangular. This entails that (I_N-A) is invertible (as an upper triangular matrix with non-zero diagonal coefficients). The low-level variables then satisfy

$$\boldsymbol{X}^{(\boldsymbol{i})}=(I_N-A)^{-1}(\boldsymbol{U}+\boldsymbol{i})$$

where

$$(I_N-A)^{-1}=\begin{bmatrix}(I_{N_0}-A_{00})^{-1}, & (I_{N_0}-A_{00})^{-1}A_{01}(I_{N_1}-A_{11})^{-1}\\0 & (I_{N_1}-A_{11})^{-1}\end{bmatrix}$$

leading to the pushforward high-level cause variable

$$\widehat{Z}_1^{(\boldsymbol{i})}=\boldsymbol{\tau}_1^\top(I_N-A)^{-1}(\boldsymbol{U}+\boldsymbol{i})=\bar{\boldsymbol{\tau}}_1^\top(I_{N_1}-A_{11})^{-1}(\boldsymbol{U}_{\pi(1)}+\boldsymbol{i}_{\pi(1)})\,.$$

Looking for the solutions satisfying $\mathcal{L}_{cons}=0$ entails that we must satisfy that for almost all $\boldsymbol{i}$, this push-forward distribution matches the learned high-level interventional distribution which satisfies

$$P^{(\omega_1(\boldsymbol{i}))}(Z_1)=P^{(0)}\left(Z_1-\boldsymbol{\omega}_1^\top(\boldsymbol{i})\right)\,.$$

According to Assum. B.2, and since we assume $\Omega=\pi(1)$, the prior $P(\boldsymbol{i}_{\pi(1)})$ has density with respect to the Lebesgue measure with support including a neighborhood of $\boldsymbol{i}=\boldsymbol{0}$, by continuity of the KL divergence, a solution making the consistency loss vanish needs to have the KL divergence term vanish for $\boldsymbol{i}=\boldsymbol{0}$ (otherwise we could find a neighborhood of $\boldsymbol{i}=\boldsymbol{0}$ such that the KL does not vanish, by continuity of the KL divergence).

This vanishing of the KL divergence entails that its terms, the two unintervened densities, are equal, such that we get

$$P^{(0)}(Z_1) = (\bar{\tau}_1^\top (I_{N_1} - A_{11})^{-1})_\# P(\boldsymbol{U}) = \mathcal{N}(\bar{\tau}_1^\top (I_N - A_{11})^{-1} \mu_{\boldsymbol{U}_1}, \bar{\tau}_1^\top (I_N - A_{11})^{-1} \Sigma_{\boldsymbol{U}_{\pi(1)}} (I_N - A_{11})^{-\top} \bar{\tau}_1) \,,$$

which entails

$$\sigma^2_{Z,1} = \bar{\tau}_1^\top (I_N - A_{11})^{-1} \Sigma_{\boldsymbol{U}_{\pi(1)}} (I_N - A_{11})^{-\top} \bar{\tau}_1 \tag{17}$$

and

$$\mu_{Z,1} = \bar{\tau}_1^\top (I_N - A_{11})^{-1} \mu_{\boldsymbol{U}_{\pi(1)}} \,. \tag{18}$$

For the same reasons, we can further match the interventional distributions in an open set included in the interior of the support of $P(\boldsymbol{i})$, such that for all $\boldsymbol{i}$ in this open set

$$\begin{aligned} &\mathcal{N}(\bar{\tau}_1^\top (I_{N_1} - A_{11})^{-1} (\mu_{\boldsymbol{U}_1} + \boldsymbol{i}), \bar{\tau}_1^\top (I_{N_1} - A_{11})^{-1} \Sigma_{\boldsymbol{U}_1} (I_{N_1} - A_{11})^{-\top} \bar{\tau}_1) \\ &\qquad = \mathcal{N}(\bar{\tau}_1^\top (I_{N_1} - A_{11})^{-1} \mu_{\boldsymbol{U}_1} + \boldsymbol{\omega}_1^\top \boldsymbol{i}, \bar{\tau}_1^\top (I_{N_1} - A_{11})^{-1} \Sigma_{\boldsymbol{U}_1} (I_{N_1} - A_{11})^{-\top} \bar{\tau}_1) \,. \end{aligned}$$

Indeed, otherwise the KL would not vanish in a neighborhood of non-zero measure and would contradict the fact that $\mathcal{L}_{cons}$ vanishes.

This implies that for all $\boldsymbol{i}$ in this open neighborhood

$$\bar{\tau}_1^\top (I_{N_1} - A_{11})^{-1} (\mu_{\boldsymbol{U}_{\pi(1)}} + \boldsymbol{i}_{\pi(1)}) = \bar{\tau}_1^\top (I_{N_1} - A_{11})^{-1} \mu_{\boldsymbol{U}_{\pi(1)}} + \boldsymbol{\omega}_1^\top \boldsymbol{i}_{\pi(1)} \,,$$

which simplifies to

$$\bar{\tau}_1^\top (I_{N_1} - A_{11})^{-1} \boldsymbol{i}_{\pi(1)} = \boldsymbol{\omega}_1^\top \boldsymbol{i}_{\pi(1)} \,.$$

Since this equality between two affine functions of $\boldsymbol{i}_{\pi(1)}$ is valid on an open set of the vector space of $\boldsymbol{i}_{\pi(1)}$, these affine functions must be equal (we can reparameterize $\boldsymbol{i}$ to show that constants must match at the new origin and the linear maps must match on a basis of the space, so they are equal). This is valid if and only if, in addition to (17), (18),

$$\boldsymbol{\omega}_1 = (I_{N_1} - A_{11})^{-\top} \bar{\tau}_1 \,, \tag{19}$$

is verified.

Alternatively, we obtain the same result by replacing Assum. B.2 by Assum. B.3. Indeed, the finite distribution over interventions imposes that the KL term inside the expectation must vanish for each of them (including the unintervened distribution). As long as the collection of finite interventions vectors forms a rank $\pi(1)$ family, we can choose a subset of $\pi(1)$ such vectors $\{\boldsymbol{i}^1_{\pi(1)}, \ldots, \boldsymbol{i}^{\pi(1)}_{\pi(1)}\}$ such that it forms a linearly independent family. It can be used to build the matrix equality

$$\bar{\tau}_1^\top (I_{N_1} - A_{11})^{-1} \left[\boldsymbol{i}^1_{\pi(1)}, \ldots, \boldsymbol{i}^{\pi(1)}_{\pi(1)}\right] = \boldsymbol{\omega}_1^\top \left[\boldsymbol{i}^1_{\pi(1)}, \ldots, \boldsymbol{i}^{\pi(1)}_{\pi(1)}\right] \tag{20}$$

where the matrix $\left[\boldsymbol{i}^1_{\pi(1)}, \ldots, \boldsymbol{i}^{\pi(1)}_{\pi(1)}\right]$ is invertible. By right-multiplying Eq. (20) by this inverse, we obtain Eq. (19) again.

We can move on to check the implication of consistency of the effect's conditional. It entails for almost all of $\boldsymbol{i}$

$$\widehat{P}^{(\boldsymbol{i})}(Y|Z) \widehat{P}^{(\boldsymbol{i})}(Z) = P^{(\omega(\boldsymbol{i}))}(Y|Z) \widehat{P}^{(\boldsymbol{i})}(Z) = P^{(0)}(Y|Z) \widehat{P}^{(\boldsymbol{i})}(Z) \,.$$

The left-hand side is obtained by using

$$(Y, Z) \sim \mathcal{N}(T\mu_{\boldsymbol{X}}, T\Sigma_{\boldsymbol{X}} T^\top) \,.$$

And the right-hand side by using

$$Y = f(Z) + R_0 \,.$$

Fitting first only the marginals of Y, we obtain necessary conditions. We have

$$\widehat{P}^{(\boldsymbol{i})}(Y) = P^{(\omega(\boldsymbol{i}))}(Y)$$

where for the left-hand side

$$Y = \bar{\tau}_0^\top (I_{N_0} - A_{00})^{-1} \boldsymbol{U}_{\pi(0)} + \bar{\tau}_0^\top (I_{N_0} - A_{00})^{-1} A_{01} (I_{N_1} - A_{11})^{-1} (\boldsymbol{U}_{\pi(1)} + \boldsymbol{i})$$

and for the right-hand side

$$Y \sim f_{\#}[P^{(\omega(\boldsymbol{i}))}(Z_1)] * P(R_0)\,.$$

As we look for linear Gaussian high-level models, we assume f is affine and parametrize it as $f(Z) = \alpha Z + \beta$. Then, under Assum. B.2, equality of marginal distributions entails the following equality for all $\boldsymbol{i}_{\pi(0)}$ in an open neighborhood of 0 (otherwise $\mathcal{L}_{rel} \leq \mathcal{L}_{cons}$ would not vanish)

$$\bar{\boldsymbol{\tau}}_0^\top (I_{N_0} - A_{00})^{-1} \boldsymbol{\mu}_{\boldsymbol{U}_0} + \bar{\boldsymbol{\tau}}_0^\top (I_{N_0} - A_{00})^{-1} A_{01} (I_{N_1} - A_{11})^{-1} (\boldsymbol{\mu}_{\boldsymbol{U}_1} + \boldsymbol{i}_{\pi(1)}) = \alpha \bar{\boldsymbol{\tau}}_1^\top (I_{N_1} - A_{11})^{-1} (\mu_{\boldsymbol{U}_1} + \boldsymbol{i}_{\pi(1)}) + \beta + \boldsymbol{\mu}_{R_0}$$

which requires (setting $\boldsymbol{i} = 0$)

$$+\beta + \boldsymbol{\mu}_{R_0} = \bar{\boldsymbol{\tau}}_0^\top (I_{N_0} - A_{00})^{-1} \boldsymbol{\mu}_{\boldsymbol{U}_{\pi(0)}}\,.$$

We can fix $\boldsymbol{\mu}_{R_0}$ to zero to avoid redundancy, such that

$$\mu_{Y|Z=0} = \beta = \bar{\boldsymbol{\tau}}_0^\top (I_{N_0} - A_{00})^{-1} \boldsymbol{\mu}_{\boldsymbol{U}_{\pi(0)}} \tag{21}$$

and consistency of non-zero shift interventions additionally entail for all $\boldsymbol{i}$ in the support of $P(\boldsymbol{i})$

$$\bar{\boldsymbol{\tau}}_0^\top (I_{N_0} - A_{00})^{-1} A_{01} (I_{N_1} - A_{11})^{-1} \boldsymbol{i}_{\pi(1)} = \alpha \bar{\boldsymbol{\tau}}_1^\top (I_{N_1} - A_{11})^{-1} \boldsymbol{i}_{\pi(1)}\,.$$

This yields (since one can always choose a linearly independent family of vectors $\boldsymbol{i}_{\pi(1)}$ within the open neighborhood of zero for which this equality holds)

$$\bar{\boldsymbol{\tau}}_0^\top (I_{N_0} - A_{00})^{-1} A_{01} (I_{N_1} - A_{11})^{-1} = \alpha \bar{\boldsymbol{\tau}}_1^\top (I_{N_1} - A_{11})^{-1}\,.$$

Then, right-multiplying by $(I_{N_1} - A_{11})$, we get

$$A_{01}^\top (I_{N_0} - A_{00})^{-\top} \bar{\boldsymbol{\tau}}_0 = \alpha \bar{\boldsymbol{\tau}}_1\,. \tag{22}$$

Similarly as above, the same conclusion can be drawn if we replace Assum. B.2 by Assum. B.3.

We can finally check that the conditional distributions are matching. Let us first compute the covariance matrix of the low-level variables.

$$cov(\boldsymbol{X}) = \begin{bmatrix} (I_{N_0} - A_{00})^{-1} & (I_{N_0} - A_{00})^{-1} A_{01} (I_{N_1} - A_{11})^{-1} \\ \mathbf{0} & (I_{N_1} - A_{11})^{-1} \end{bmatrix} \Sigma_U \begin{bmatrix} (I_{N_0} - A_{00})^{-\top} & \mathbf{0} \\ (I_{N_1} - A_{11})^{-\top} A_{01}^\top (I_{N_0} - A_{00})^{-\top} & (I_{N_1} - A_{11})^{-\top} \end{bmatrix}.$$

Because the exogenous covariance is diagonal, we get

$$= \left[\begin{array}{cc} (I_{N_0} - A_{00})^{-1} A_{01} (I_{N_1} - A_{11})^{-1} \Sigma_{\pi(1)} (I_{N_1} - A_{11})^{-\top} A_{01}^\top (I_{N_0} - A_{00})^{-\top} & (I_{N_0} - A_{00})^{-1} A_{01} (I_{N_1} - A_{11})^{-1} \Sigma_{\pi(1)} (I_{N_1} - A_{11})^{-} \\ + (I_{N_0} - A_{00})^{-1} \Sigma_{\pi(0)} (I_{N_0} - A_{00})^{-\top} & \\ \hline (I_{N_1} - A_{11})^{-1} \Sigma_{\pi(1)} (I_{N_1} - A_{11})^{-\top} A_{01}^\top (I_{N_0} - A_{00})^{-\top} & (I_{N_1} - A_{11})^{-1} \Sigma_{\pi(1)} (I_{N_1} - A_{11})^{-\top} \end{array}\right.$$

Then,

$$cov((\widehat{Y}, \widehat{Z}_1)) = T cov(\boldsymbol{X}) T^\top$$

and

$$\mu_{\widehat{Y}|\hat{z}_1} = \mu_{\widehat{Y}} + \bar{\boldsymbol{\tau}}_0^\top (I_{N_0} - A_{00})^{-1} A_{01} (I_{N_1} - A_{11})^{-1} \Sigma_{\pi(1)} (I_{N_1} - A_{11})^{-\top} \bar{\boldsymbol{\tau}}_1 \left(\bar{\boldsymbol{\tau}}_1^\top (I_{N_1} - A_{11})^{-1} \Sigma_{\pi(1)} (I_{N_1} - A_{11})^{-\top} \bar{\boldsymbol{\tau}}_1 \right)^{-1} (\hat{z}_1 - \mu_{Z_1})$$

Thus using the above equation

$$\mu_{\widehat{Y}|\hat{z}_1} = \mu_{\widehat{Y}} + \alpha \left(\hat{z}_1 - \mu_{Z_1} \right) = \bar{\boldsymbol{\tau}}_0^\top (I_{N_0} - A_{00})^{-1} \boldsymbol{\mu}_{\boldsymbol{U}_0} + \bar{\boldsymbol{\tau}}_0^\top (I_{N_0} - A_{00})^{-1} A_{01} (I_{N_1} - A_{11})^{-1} \boldsymbol{\mu}_{\pi(1)} + \alpha \hat{z}_1 - \alpha \bar{\boldsymbol{\tau}}_1^\top (I_{N_1} - A_{11})^{-1} \boldsymbol{\mu}_{\pi(1)}\,,$$

which further simplifies with the same equation to

$$\mu_{\widehat{Y}|\hat{z}_1} = \mu_{\widehat{Y}} + \alpha \left(\hat{z}_1 - \mu_{Z_1} \right) = \bar{\boldsymbol{\tau}}_0^\top (I_{N_0} - A_{00})^{-1} \boldsymbol{\mu}_{\boldsymbol{U}_0} + \alpha \hat{z}_1\,.$$

Moreover,

$$\begin{aligned} \mathrm{var}(\widehat{Y}|\hat{z}_1) = \sigma_{\widehat{Y}}^2 - \boldsymbol{\tau}_0^\top (I_{N_0} - A_{00})^{-1} A_{01} (I_{N_1} - A_{11})^{-1} \Sigma_{\pi(1)} (I_{N_1} - A_{11})^{-\top} \bar{\boldsymbol{\tau}}_1 \left(\bar{\boldsymbol{\tau}}_1^\top (I_{N_1} - A_{11})^{-1} \Sigma_{\pi(1)} (I_{N_1} - A_{11})^{-\top} \bar{\boldsymbol{\tau}}_1 \right)^{-1} \\ \bar{\boldsymbol{\tau}}_1^\top (I_{N_1} - A_{11})^{-1} \Sigma_{\pi(1)} (I_{N_1} - A_{11})^{-\top} A_{01}^\top (I_{N_0} - A_{00})^{-\top} \bar{\boldsymbol{\tau}}_0 \end{aligned}$$

again, using the above equation this leads to the simplification

$$\begin{aligned}\mathrm{var}(\widehat{Y}|\hat{z}_1) &= \sigma^2_{\widehat{Y}} - \alpha\bar{\boldsymbol{\tau}}_1^\top (I_{N_1} - A_{11})^{-1}\Sigma_{\pi(1)}(I_{N_1} - A_{11})^{-\top}A_{01}^\top (I_{N_0} - A_{00})^{-\top}\bar{\boldsymbol{\tau}}_0 \\ &= \sigma^2_{\widehat{Y}} - \alpha\bar{\boldsymbol{\tau}}_1^\top (I_{N_1} - A_{11})^{-1}\Sigma_{\pi(1)}(I_{N_1} - A_{11})^{-\top}\bar{\boldsymbol{\tau}}_1\alpha \\ &= \bar{\boldsymbol{\tau}}_0^\top (I_{N_0} - A_{00})^{-1}\Sigma_{\pi(0)}(I_{N_0} - A_{00})^{-\top}\bar{\boldsymbol{\tau}}_0\,.\end{aligned}$$

For the high-level distribution we get

$$P(Y|z) = \mathcal{N}(\alpha z + \beta, \sigma^2_{Y|Z})$$

where we can identify all parameters with the above equations.

Now assume there exists an n-causes solution ($n > 1$) $(\tau'_k, \omega'_k, \pi')_{k=1..n}$ such that the loss vanishes, but no such solution for $n+1$ causes.

Step 1: properties of exact n-cause solutions

Then it can be linked to the 1-cause solution, which is guaranteed to exist according to our set of assumptions. Indeed, the existence of the n-cause solution implies that the distribution of the low level causal model satisfies

$$\widehat{P}^{(\boldsymbol{i})}(Y|\boldsymbol{Z} = \boldsymbol{z}) \sim \mathcal{N}\left(\sum_k \alpha_k z_k + \beta, \sigma^2_{Y|Z}\right)$$

and

$$\widehat{P}^{(\omega(\boldsymbol{i}))}(\boldsymbol{Z}) \sim \prod_k \widehat{P}^{(0)}(Z_k - \omega_k(\boldsymbol{i}_k))$$

which can be rewritten. If we define the aggregates cause $\tilde{Z} = \sum_k \alpha_k z_k = \sum_k \alpha_k \tau_k(x)$, then we can rewrite the above model as

$$\widehat{P}^{(\boldsymbol{i})}(Y|\tilde{Z} = \tilde{z}) \sim \mathcal{N}\left(\tilde{z} + \beta, \sigma^2_{Y|Z}\right)$$

and

$$\widehat{P}^{(\omega(\boldsymbol{i}))}(\tilde{Z}) \sim \prod_k \widehat{P}^{(0)}(\tilde{Z} - \sum_k \omega_k(\boldsymbol{i}_k))$$

which implies that the abstraction consists of concatenating the τ_k with multiplicative coefficient α_k.

Moreover, the interventional consistency of the n-causes, which do not influence each other according to the assumed high-level causal graph, entails that any low-level intervention $\boldsymbol{i}$ affects only high-level variable Z_k through is components in $\pi(k)$.

Consistency implies (using the above 1 cause solution proof)

$$\begin{bmatrix}\bar{\boldsymbol{\tau}}_1^\top, & \dots, & 0 \\ \vdots, & \ddots, & \vdots \\ 0, & \dots, & \bar{\boldsymbol{\tau}}_n^\top\end{bmatrix}(I_{\sum_k N_k} - A_{1..n,1..n})^{-1}\boldsymbol{i}_\Omega = \begin{bmatrix}\boldsymbol{\omega}_1^\top \boldsymbol{i}_{\pi_1} \\ \vdots \\ \boldsymbol{\omega}_n^\top \boldsymbol{i}_{\pi_n}\end{bmatrix}.$$

this implies that $\bar{\boldsymbol{\tau}}_k^\top (I_{\sum_k N_k} - A_{1..n,1..n})^{-1}_{kj} = 0$ for all $j \in \pi(l), l \neq k$. Because the non-vanishing coefficients of $\bar{\tau_k}$ reflect the influences along the causal pathway from nodes of $\pi(k)$ to Y, the above entails that $(I_{\sum_k N_k} - A_{1..n,1..n})^{-1}_{kj}$ must vanish on the support of $\bar{\tau_k}$.

We thus deduce that any off-diagonal block element of $(I_{\sum_k N_k} - A_{1..n,1..n})^{-1}$ whose row component belongs to the support of any τ_k and whose column component belongs to the support of any ω_k must vanish. Indeed, otherwise the causes would influence each other. This means essentially that any node influencing the target must not influence any node in another group with the same property.

Step 2: identifiability of the n-cause solution Now consider a second n-cause solution which we denote with a prime: $(\tau'_k, \omega'_k, \pi')_{k=1..n}$.

If $\pi' = \pi$, then identifiability of the corresponding 1-cause solution implies that each τ'_k is identified with each τ_k up to a multiplicative constant, because they are equal to the one cause tau on the same support.

Otherwise, $\pi' \neq \pi$ and we may have overlap between supports of omegas and taus. If there is actual overlap, then the vanishing of the off-diagonal block coefficients on the overlapping support entails that one can design a separate group of variables that can be turned in an $n+1$ independent cause, which contradicts the assumptions.

□

D.3 Proof of Proposition D.2

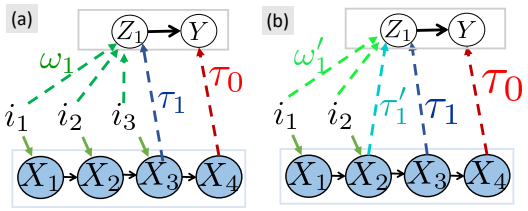


Figure 3: **TCR solutions on a chain graph.** Arrows indicate the non-zero coefficients of each map. (a) Unique solution τ_1 when interventions are on all nodes but the target. (b) Two solutions τ_1 and τ'_1 when interventions are on the first two nodes only.

Proposition D.2. *Consider the setting of Prop. B.4 with the exception that $\Omega \subsetneq \pi(1)$ such that there is now a non-empty subset $S = \pi(1) \setminus \Omega$, such that $\boldsymbol{X}_\Omega \to \boldsymbol{X}_S \to \boldsymbol{X}_{\pi(0)}$. Then there exist at least two different linear 1D TCR such that $\mathcal{L}_{cons} = 0$.*

This result can also be illustrated with a chain graph, as shown in Fig. 3(b). If the parent node X_3 of $Y = X_4$ is unintervened, then one may choose either $Z_1 = X_2$ or $Z_1 = X_3$ (matching the solution of Fig. 3(a)) to minimize $\mathcal{L}_{cons}$. This is because both variables are equivalently mediating all performed interventions to X_4. Note that each choice has it own benefit: $Z_1 = X_3$, as a direct parent of Y, is a better statistical predictor of the value of Y. However, if we focus on causal interpretability of the high-level representation, $Z_1 = X_2$ is preferable because it is one of the variables intervened on at the low-level as enforced by the prior $P(\boldsymbol{i})$, and such that it will be associated to a non-zero weight in $\boldsymbol{\omega}_1$ for any solution satisfying $\mathcal{L}_{cons} = 0$.

Proof. The low-level model follows the following SCM, with $P(\boldsymbol{i})$ non-trivial

$$\boldsymbol{X} \coloneqq A\boldsymbol{X} + \boldsymbol{U} + \boldsymbol{i}\,, \quad U_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$$

such that $\boldsymbol{X}$, A and P take the block forms

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_{\pi(0)} \\ \boldsymbol{X}_S \\ \boldsymbol{X}_\Omega \end{bmatrix}, \quad A = \begin{bmatrix} A_{00} & A_{0S} & \mathbf{0} \\ \mathbf{0} & A_{SS} & A_{S\Omega} \\ \mathbf{0} & \mathbf{0} & A_{\Omega\Omega} \end{bmatrix},$$

with $\pi(0)$ of size N_0, and $\pi(1)$ of size $N_1 = N - N_0$ and S of size s. Then we know from the Proposition B.4 that there is already a valid solution using π as alignment. The only difference is that variables in S are unintervened, which does not affect the ability of the solution to achieve $\mathcal{L}_{cons} = 0$. That solution would be compatible with interventions on S, but since S is unintervened, we do not have uniqueness guarantees for this choice of π.

Alternatively, if we choose $\pi'(0) = \pi(0) \cup S$ and $\pi'(1) = \pi(1) \setminus S = \Omega$, then, we can again apply Propostion D.2, and see that it provides a different solution with this alignment, which is compatible with the given problem (constructive transformation with constraint on the mapping τ_0). Importantly, the key indeterminacy is for the map τ_1, which will either put all its weight on elements in S (direct parents of $\pi(0)$), or alternatively, put all its weights on elements in Ω. There is an additional, but trivial, indeterminacy for the map ω_1: indeed, since X_S is unintervened (as part of $\pi(0)$), the weights in $\boldsymbol{\omega}_1$ associated to these coefficients may take arbitrary values (since their associated component in $\boldsymbol{i}$ remains zero). We do not consider these trivial indeterminacies (which do not affect the mapping ω_1 on its domain, i.e. the support of the prior $P(\boldsymbol{i})$) by forcing the weights of $\boldsymbol{\omega}_1$ associated to unintervened variables to zero. □

E Additional Theory

E.1 Reparametrizations of Reductions

In order to study invariance properties of TCR, we define transformations compatible with a class of reductions. Let $\rho : \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_n \to \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_n$ be a continuous invertible transformation of the n-dimensional high-level cause vector. Then the transformation

$$\tilde{\rho} : \begin{bmatrix} Y \\ \boldsymbol{Z} \end{bmatrix} \mapsto \begin{bmatrix} Y \\ \rho(\boldsymbol{Z}) \end{bmatrix}$$

is also continuous invertible. Among this class of transformations, we define an invertible reparametrization of a TCR as follows.

Definition E.1. *An invertible reparametrization of a reduction for the class $\mathcal{T}$ of τ-maps and the class $\{\mathcal{H}_{\boldsymbol{\gamma}}\}_{\boldsymbol{\gamma}\in\Gamma}$ satisfies the following properties.*

- *it is* compatible *with the class of τ-maps as follows: for any map $\tau \in \mathcal{T}$, we have $\tilde{\rho} \circ \tau \in \mathcal{T}$,*
- *it is* compatible *with the high-level model class $\{\mathcal{H}_{\boldsymbol{\gamma}}\}$ as follows: for any model parameter $\boldsymbol{\gamma}$, the unintervened and intervened distributions $P_{\mathcal{H},\boldsymbol{\gamma}}(Y, \boldsymbol{Z})$ are such that there exist a parameter $\boldsymbol{\gamma}'$ and a map between high-level interventions $\psi : \mathcal{J} \to \mathcal{J}$ such that the joint distributions of the transformed variables $(Y, \rho(\boldsymbol{Z}))$ is compatible with unintervened and intervened distributions of $\mathcal{H}_{\boldsymbol{\gamma}'}$, in the sense that*

$$\tilde{\rho}_{\#}[P^{(\boldsymbol{j})}_{\mathcal{H},\boldsymbol{\gamma}}(Y, \boldsymbol{Z})] = P^{(\psi(\boldsymbol{j}))}_{\mathcal{H},\boldsymbol{\gamma}'}(Y, \boldsymbol{Z})\,.$$

E.2 The Case of a Single Target Low-Level Variable

Whenever $\pi(0)$ is a singleton, τ_0 is univariate and the target Y essentially corresponds (up to trivial rescaling) to a single low-level variable. We elaborate on the interpretation of Proposition B.4 in this context.

Let us set $\bar{\boldsymbol{\tau}}_0 = 1$ and fix the target index such that $\pi(0) = \{N\}$ without loss of generality. Then the DAG constraints entail $A_{00} = 0$ and the structural equations take the form

$$\boldsymbol{X}_{\pi(1)} \coloneqq A_{11}\boldsymbol{X}_{\pi(1)} + \boldsymbol{U}_{\pi(1)} + \boldsymbol{i}\,, \quad U_k \sim \mathcal{N}(\mu_k, \sigma_k^2) \tag{23}$$

$$Y \coloneqq \boldsymbol{a}_{01}^{\top}\boldsymbol{X}_{\pi(1)} + U_N \tag{24}$$

where $\boldsymbol{a}_{01}$ is a column vector of coefficients of the low-level mechanism linking the target Y to its causes in $\pi(0)$. Then the unique linear 1D TCR, up to a multiplicative constant, making the consistency loss vanish is given by

$$\bar{\boldsymbol{\tau}}_1 = \boldsymbol{a}_{01} \tag{25}$$

$$\text{and} \quad \bar{\boldsymbol{\omega}}_1 = (I_{N-1} - A_{11})^{-\top}\bar{\boldsymbol{\tau}}_1 = (I_{N-1} - A_{11})^{-\top}\boldsymbol{a}_{01}\,. \tag{26}$$

This solution is easily interpretable: $\bar{\boldsymbol{\tau}}_1$ identifies the ground truth mechanism linking $\boldsymbol{X}_{\pi(0)}$ to the target, while $\bar{\boldsymbol{\omega}}_1$ traces the contribution of interventions on each endogenous variable to the target. Indeed, this contribution is given by the "reduced form" map between exogenous values and endogenous values (see proof of Proposition B.4 for more insights)

$$\boldsymbol{i} \mapsto (I_{N-1} - A_{11})^{-1}\boldsymbol{i}\,,$$

and by composing this mapping with mechanism $\boldsymbol{a}_{01}$ we get the (shift) influence of interventions on the target

$$\boldsymbol{i} \mapsto \boldsymbol{a}_{01}^{\top}(I_{N-1} - A_{11})^{-1}\boldsymbol{i} = \bar{\boldsymbol{\omega}}_1^{\top}\boldsymbol{i}\,.$$

The mismatch between $\bar{\boldsymbol{\omega}}_1$ and $\bar{\boldsymbol{\tau}}_1$ is due to the internal causal structure of the submodel described by eq. (23). Indeed, if there are no causal links within this subsystem, A_{11} is a zeros matrix and

$$\bar{\boldsymbol{\omega}}_1 = (I_{N-1})^{-\top}\bar{\boldsymbol{\tau}}_1 = \bar{\boldsymbol{\tau}}_1 = \boldsymbol{a}_{01}\,,$$

otherwise, the two maps will be different. The discrepancy between the vectors thus reflects the fact that the causal explanation links high-level endogenous variables and interventions on them by potentially complex low-level interactions that do not necessarily have a simple high-level interpretation. This justifies regularizing the consistency loss with an homogeneity loss in order to focus on explanations that exhibit congruent τ and ω maps.

E.3 The Case of Linear Chain SCMs

In the case of a chain SCM

$$\boldsymbol{X}_1 \to \cdots \to \boldsymbol{X}_{N-1} \to \boldsymbol{X}_N = Y$$

the above linear setting gets the additional constraints (using a causal ordering of the variables) that the target's mechanism is sparse

$$\boldsymbol{a}_{01}^\top = [0, \ldots, 0, a_N]$$

and the structure matrix of $\boldsymbol{X}_{\pi(1)}$ is subdiagonal

$$A_{11} = \begin{bmatrix} 0 & 0 & \ldots & 0 & 0 \\ a_2 & 0 & \ldots & 0 & 0 \\ 0 & a_3 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & a_{N-1} & 0 \end{bmatrix}$$

and as a consequence, the solution writes

$$\bar{\boldsymbol{\tau}}_1 = [0, \ldots, 0, a_{N-1}]^\top \tag{27}$$

$$\text{and} \quad \bar{\boldsymbol{\omega}}_1 = (I_{N-1} - A_{11})^{-\top} \bar{\boldsymbol{\tau}}_1 = \begin{bmatrix} a_2.a_3.\ldots.a_{N-1} \\ \vdots \\ a_{N-2}a_{N-1} \\ a_{N-1} \end{bmatrix}. \tag{28}$$

This solution is in line with our experimental results:

- $\bar{\boldsymbol{\tau}}_1$ has all its weight on the parent of the target.
- $\bar{\boldsymbol{\omega}}_1$ has a non-sparse distribution over the chains, decaying in the upstrean direction. This reflects that structure coefficients of A_{11} are selected with absolute value inferior to one, such that the influence of ancestor nodes on the target decays with their distance to it on the graph.

Transposing the chain example to the case of Proposition D.2, we can take the case were the direct parent X_{N-1} of the target is left unintervened. In such a case, $\bar{\boldsymbol{\tau}}_1$ may put its weight on both X_{N-1} and its direct parent X_{N-2}, Proposition B.4 provides two example solutions for different choices of $\pi(1)$, including or excluding X_{N-1}. In the most extreme case of dissimilarity between $\boldsymbol{\tau}_1$ and $\boldsymbol{\omega}_1$, solution including X_{N-1} in $\pi(1)$ puts all $\boldsymbol{\tau}_1$'s weight on X_{N-1}, while $\boldsymbol{\omega}_1$ has no weight on it (because it is unintervened). As a consequence, $\boldsymbol{\omega}_1$ and $\boldsymbol{\tau}_1$ are orthogonal and the associated homogeneity loss vanishes. In contrast, the unique solution excluding X_{N-1} from $\pi(1)$ have a larger cosine similarity and will thus be preferred by the homogenity-regularized loss.

F Algorithm Details

F.1 Gaussian Consistency Loss

As the KL divergence is hard to estimate in the non-parametric setting, we make a Gaussian approximation of this loss to get an analytical, differentiable expression. Using the general formula for two n-dimensional Gaussian densities P and Q

$$KL(P||Q) = \frac{1}{2}\left[(\mu_Q - \mu_P)^\top \Sigma_Q^{-1}(\mu_Q - \mu_P) + \text{tr}(\Sigma_Q^{-1}\Sigma_P) - \log\frac{|\Sigma_P|}{|\Sigma_Q|} - n\right].$$

Parameters of the reduction are $\boldsymbol{\tau}_k, \boldsymbol{\mu}_Z, \mu_{Y|Z}, f : z \to f(z), \boldsymbol{\omega}_k$ with

$$\begin{aligned} Z^{(\omega(\boldsymbol{i}))} &\sim P(z) = \mathcal{N}(\boldsymbol{\mu}_Z + W\boldsymbol{i}, \Sigma_{\boldsymbol{Z}}), \text{ with } W = [\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_n]^\top \text{ and } \Sigma_{\boldsymbol{Z}} = \text{diag}(\sigma^2_{Z,1}, ..., \sigma^2_{Z,n})) \\ Y^{(\omega(\boldsymbol{i}))}|\boldsymbol{z} &\sim P(Y|z) = \mathcal{N}(f(\boldsymbol{z}), \sigma^2_{Y|\boldsymbol{Z}}), \\ \hat{Z}^{(\boldsymbol{i})} &= [\boldsymbol{\tau}_1, ..., \boldsymbol{\tau}_n]^\top \boldsymbol{X}^{(\boldsymbol{i})} = T\boldsymbol{X}^{(\boldsymbol{i})}, \\ \hat{Y}^{(\boldsymbol{i})} &= \boldsymbol{\tau}_0^\top \boldsymbol{X}^{(\boldsymbol{i})}. \end{aligned}$$

Moreover, we estimate the second order properties of the simulator distribution for each intervention $\boldsymbol{i}$

$$
\begin{aligned}
\hat{\boldsymbol{\mu}}_{\boldsymbol{X}}^{(\boldsymbol{i})} &= \langle \boldsymbol{X}^{(\boldsymbol{i})} \rangle \,, \\
\hat{\Sigma}_{\boldsymbol{X}}^{(\boldsymbol{i})} &= \left\langle \left(\boldsymbol{X}^{(\boldsymbol{i})} - \hat{\boldsymbol{\mu}}_{\boldsymbol{X}}^{(\boldsymbol{i})} \right)^{\top} \left(\boldsymbol{X}^{(\boldsymbol{i})} - \hat{\boldsymbol{\mu}}_{\boldsymbol{X}}^{(\boldsymbol{i})} \right) \right\rangle \,, \\
\hat{\boldsymbol{\mu}}_{\boldsymbol{Z}}^{(\boldsymbol{i})} &= \langle \hat{Z}^{(\boldsymbol{i})} \rangle = T \hat{\boldsymbol{\mu}}_{\boldsymbol{X}}^{(\boldsymbol{i})} \,, \\
\hat{\mu}_{Y}^{(\boldsymbol{i})} &= \langle \hat{Y}^{(\boldsymbol{i})} \rangle = \boldsymbol{\tau}_0^{\top} \hat{\boldsymbol{\mu}}_{\boldsymbol{X}}^{(\boldsymbol{i})} \,, \\
\widehat{\Sigma}_{\boldsymbol{Z}}^{(\boldsymbol{i})} &= \left\langle \left(\widehat{\boldsymbol{Z}}^{(\boldsymbol{i})} - \hat{\boldsymbol{\mu}}_{Z}^{(\boldsymbol{i})} \right) \left(\widehat{\boldsymbol{Z}}^{(\boldsymbol{i})} - \hat{\boldsymbol{\mu}}_{Z}^{(\boldsymbol{i})} \right)^{\top} \right\rangle = T \hat{\Sigma}_{\boldsymbol{X}}^{(\boldsymbol{i})} T^{\top} \,, \\
\widehat{\sigma^2}_{Z,k}^{(\boldsymbol{i})} &= \left(\widehat{\Sigma}_{\boldsymbol{Z}}^{(\boldsymbol{i})} \right)_{k,k} = \left\langle \left(\widehat{Z}_k^{(\boldsymbol{i})} - \hat{\mu}_{Z,k}^{(\boldsymbol{i})} \right)^2 \right\rangle = \boldsymbol{\tau}_k^{\top} \hat{\Sigma}_{\boldsymbol{X}}^{(\boldsymbol{i})} \boldsymbol{\tau}_k \,, \\
\widehat{\sigma^2}_{Y}^{(\boldsymbol{i})} &= \left\langle \left(\hat{Y}^{(\boldsymbol{i})} - \hat{\mu}_{Y}^{(\boldsymbol{i})} \right)^2 \right\rangle = \boldsymbol{\tau}_0^{\top} \hat{\Sigma}_{\boldsymbol{X}}^{(\boldsymbol{i})} \boldsymbol{\tau}_0 \,, \\
\widehat{\boldsymbol{c}}_{\boldsymbol{Z}Y}^{(\boldsymbol{i})} &= \left\langle \left(\hat{Y}^{(\boldsymbol{i})} - \hat{\mu}_{Y}^{(\boldsymbol{i})} \right) \left(\hat{\boldsymbol{Z}}^{(\boldsymbol{i})} - \hat{\boldsymbol{\mu}}_{Z}^{(\boldsymbol{i})} \right) \right\rangle = T \hat{\Sigma}_{\boldsymbol{X}}^{(\boldsymbol{i})} \boldsymbol{\tau}_0 \,,
\end{aligned}
$$

where $\langle \cdot \rangle$ denotes the empirical average. Using the KL between Gaussian variables, we can rewrite the consistency loss as

$$
\begin{aligned}
\mathcal{L}_{cons} &= \mathbb{E}_{\boldsymbol{i} \sim p(\boldsymbol{i})} \left[KL_z(\hat{P}^{(\boldsymbol{i})}(z) | P^{(\hat{\omega}(\boldsymbol{i}))}(z)) \right] + \mathbb{E}_{z \sim \hat{P}^{(\boldsymbol{i})}(\boldsymbol{Z})} \left[KL_Y \left(\hat{P}^{(\boldsymbol{i})} \left(Y | \boldsymbol{Z} = z \right) || P^{(0)} \left(Y | \boldsymbol{Z} = z \right) \right) \right] \\
&= \frac{1}{2} \mathbb{E}_{\boldsymbol{i} \sim p(\boldsymbol{i})} \left[\sum_k \left(\frac{(\mu_{Z,k} + \boldsymbol{\omega}_k^{\top} \boldsymbol{i} - \hat{\mu}_{Z,k}^{(\boldsymbol{i})})^2}{\sigma_{Z,k}^2} + \frac{\widehat{\sigma^2}_{Z,k}^{(\boldsymbol{i})}}{\sigma_{Z,k}^2} \right) - \ln \left(\frac{|\widehat{\Sigma}_{\boldsymbol{Z}}^{(\boldsymbol{i})}|}{\prod_k \sigma_{Z,k}^2} \right) - n \right] \\
&\quad + \frac{1}{2} \mathbb{E}_{\boldsymbol{i} \sim p(\boldsymbol{i}), z \sim \hat{P}^{(\boldsymbol{i})}(Z)} \left[\frac{\left(f(z) - \hat{\mu}_Y^{(\boldsymbol{i})} - \left(\widehat{\boldsymbol{c}}_{\boldsymbol{Z}Y}^{(\boldsymbol{i})} \right)^{\top} \left(\widehat{\Sigma}_{\boldsymbol{Z}}^{(\boldsymbol{i})} \right)^{-1} (z - \hat{\boldsymbol{\mu}}_Z^{(\boldsymbol{i})}) \right)^2}{\sigma_{Y|Z}^2} \right. \\
&\quad \left. + \frac{\widehat{\sigma^2}_Y^{(\boldsymbol{i})} - \left(\widehat{\boldsymbol{c}}_{\boldsymbol{Z}Y}^{(\boldsymbol{i})} \right)^{\top} \left(\widehat{\Sigma}_{\boldsymbol{Z}}^{(\boldsymbol{i})} \right)^{-1} \widehat{\boldsymbol{c}}_{\boldsymbol{Z}Y}^{(\boldsymbol{i})}}{\sigma_{Y|Z}^2} - \ln \left(\frac{\widehat{\sigma^2}_Y^{(\boldsymbol{i})} - \left(\widehat{\boldsymbol{c}}_{\boldsymbol{Z}Y}^{(\boldsymbol{i})} \right)^{\top} \left(\widehat{\Sigma}_{\boldsymbol{Z}}^{(\boldsymbol{i})} \right)^{-1} \widehat{\boldsymbol{c}}_{\boldsymbol{Z}Y}^{(\boldsymbol{i})}}{\sigma_{Y|Z}^2} \right) - 1 \right] . \qquad (29)
\end{aligned}
$$

G Experiments

G.1 Additional Results

G.1.1 Toy Models: Linear Gaussian Low-Level Causal Model

Linear low-level causal models. We first test TCR by sampling from a linear Gaussian low-level model, rather than a simulation. We construct linear models of the form shown in Proposition B.4 by drawing the non-zero entries in the adjacency matrix uniformly from the interval $[-1, 1]$. We learn a targeted causal reduction with two high-level variables: the target Y and its single cause Z. Fig. 4 compares the learned $\boldsymbol{\tau}_1$ and $\boldsymbol{\omega}_1$ to the analytical solutions (10) and (11). We observe that, for these low-level models meeting the linear Gaussian assumption in Section B, the learning algorithm converges to the global optimum.

Two-branch model. To investigate the behavior of TCR with multiple high-level variables, we consider a low-level model with two branch causal structure (Fig. 5). With regularization for overlap (12) and balancing (13), the learned high-level variables correspond to the two branches. Comprehensive experimental details are given in App. G.2.

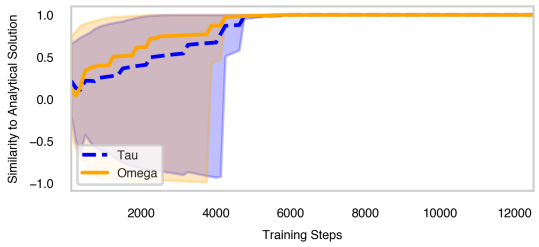


Figure 4: **Comparison between learned and analytical solutions 1-cause TCR** Average cosine similarity to the analytical solutions over 20 runs. Each run corresponds to one draw of adjacency matrix parameters. The shaded areas correspond to the range between the minimum and maximum values.

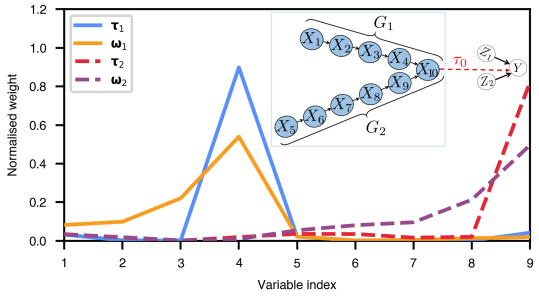


Figure 5: **Two-branch linear model.** Learned τ- and ω parameters for a TCR with two high-level variables for a linear Gaussian low-level model with N=10. The solid lines show the parameters for Z_1 and the dashed lines those for Z_2. The parameters are averaged over 20 runs where each run corresponds to one draw of adjacency matrix parameters. The inlay shows the causal structure of the low-level model, where two groups of variables G_1 and G_2 form two independent chains causing the target X_{10}=Y.

G.1.2 Spring-Mass System without Regularization

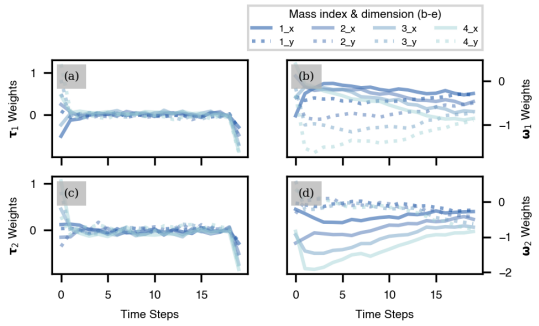


Figure 6: **Spring-mass system experiment without regularization.** Same experimental setup as described in Sec. 4 and App. G.4 with the regularization turned off, *i.e.* $\eta_{\text{ov}} = \eta_{\text{bal}} = 0$. The learned high-level mechanism is $f(\mathbf{Z}) \approx -0.180Z_1 + 0.125Z_2$.

When running the TCR algorithm without regularization, it cannot be ensured that the found solutions correspond to different properties of the low-level system, as shown in Fig. 6. There is significant mixing among the high-level variables, in particular the velocity in x-direction of the masses towards the end of the simulation appears in both high-level variables.

G.1.3 Double Well

For a simulation based on an ODE system, we learn a targeted reduction of a ball moving in a double well potential under linear friction, as shown in Fig. 7. The state vector $\boldsymbol{X}$ encodes the x-position and velocity in x-direction of the ball at each time steps of the simulation. As shift-interventions, we apply small random shifts of the ball's velocity at each simulation time step, mimicking an applied external

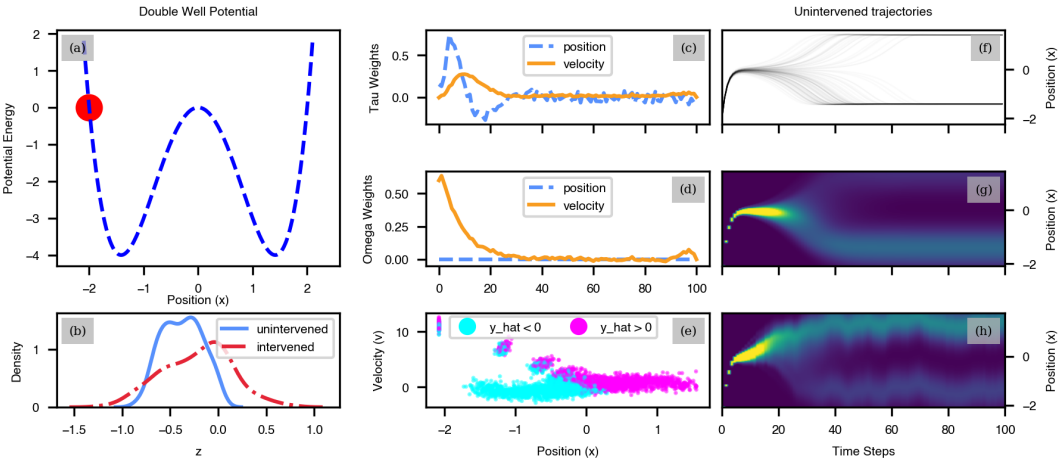


Figure 7: **Double well experiment.** (a) Depicts the experimental setup with a ball moving in a double well potential subject to linear friction. (b) Displays the pushforward density of the high-level cause for the two settings: one where no intervention is applied (unintervened), and the other with an applied shift intervention. (c) and (d) represent the learned parameters, τ and ω, respectively. The learned high-level mechanism is $f(Z_1) \approx 1.37Z_1 + 0.45$ (e) Shows samples in phase space (position vs. velocity) for the first 20 time points and whether the high-level model predicts the ball to end up in the right (pink) or right well (turquoise). (f) and (g) show samples from the unintervened setting and the corresponding estimated density. (h) Depicts the estimated density for one intervened setting.

force. Initially, the ball starts on the left-hand side of the potential and starts oscillating. Since the ball experiences friction, it ends up in either the left or right minimum of the potential. The friction is relatively strong, such that, depending on the initial conditions and applied shift interventions, the ball either stays in the left well or crosses the middle hump once and stays in the right well (see Fig. 7(f)). We learn a simple TCR with a single cause Z that explains the target Y. Further details about the nonlinear ODE system and training are given in App. G.3.

The learned TCR parameters are shown in Fig. 7(c, d). The $\boldsymbol{\tau}_1$ and $\boldsymbol{\omega}_1$ parameters for velocity are such that the larger the velocity is to the right, the higher Z and therefore the higher the predicted target Y, where positive Y correspond to the right well and negative to the left. Similarly, for the position parameter: the more negative the position just before the critical point of the ball crossing the hump, the higher the probability of predicting to stay in the left well. This corresponds to the correct dynamics of the system and also identifies the main drivers that influence the outcome Y. Fig. 7(e) shows how the learned TCR separates the phase space into simulations with enough momentum to the right to make it over the hump (pink) and those without (turquoise).

Note that TCR does not focus on the part of $\boldsymbol{X}$ which best predicts the final state of the system—like the position just before the end of the simulation. It rather highlights the variables which have the most impact on the target when they are intervened on, emphasizing the decisive time span when the ball either crosses the middle hump or stays in the left well.

G.1.4 Grouped Spring-Mass System

We simulate two groups of four masses as shown in Fig. 8(a). In contrast to the experiment shown in Sec. 4, all masses have equal weight and the target is the center of mass velocity in x-direction at the end of the simulation. The data and training parameters are summarized in Table 3.

Since the only interactions between masses are mediated by the springs, as described in App. G.4, the two groups of masses do not influence each other and are thus fully independent. The learned TCR identifies the two groups of masses as the two independent causes of the target. This is reflected in the parameters shown in Fig. 8 (b-e), where high-level variable Z_1 is predominantly influenced by the behavior of the second group (yellow) and variable Z_2 by the first group (blue). Furthermore, we observe that the y-component of the velocity, which is irrelevant for the target here, is ignored by the TCR and filtered out.

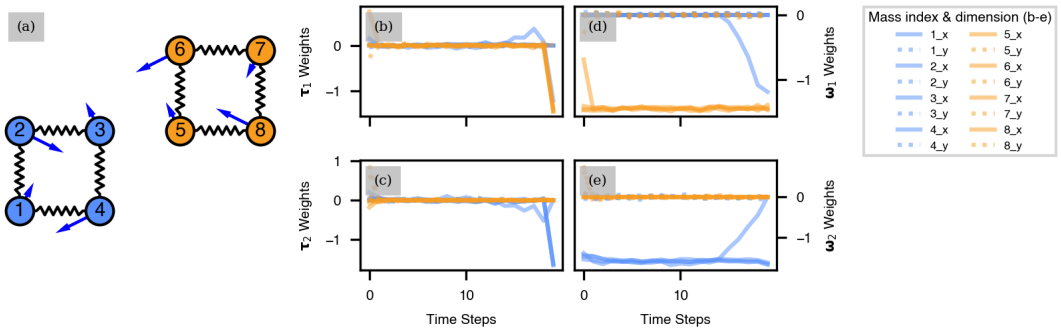


Figure 8: **Spring-mass system experiment with two groups of masses.** (a) Simulated system of eight point masses with equal weights connected by springs in two groups of 4 and with random initial velocity (blue arrows). In contrast to the experiment shown in Sec. 4, target of the simulation is the center of mass speed in x-direction. (b-e) Learned $\boldsymbol{\tau}$- and $\boldsymbol{\omega}$-weights corresponding to velocity components in x- and y-direction for a TCR with two high-level variables. The learned high-level mechanism is $f(\boldsymbol{Z}) \approx -0.0866Z_1 - 0.0782Z_2$.

G.2 Experimental Details: Linear Experiments

Parameters	**Linear** (Fig. 4)	**Two Branch** (Fig. 5)
learning rate λ	10^{-3}	10^{-3}
learning rate scheduler	-	cosine annealing
No. repeated train. runs per seed	1	10
simulation paths n_{sim}	$10,000$	$10,000$
training epochs N_{ite}	100	600
simulation batch size B	128	128
intervention batch size B_i	64	512
overlap reg. η_{ov} (12)	0	0.1
balancing reg. η_{bal} (13)	0	10^{-3}

Table 1: **Experimental parameters and settings for the linear Gaussian experiments.**

Sampling linear Gaussian low-level models For the adjacency matrix, we sample all non-zero entries uniformly in the interval $[-1, 1]$. For general adjacency matrices, the lower triangular elements of the adjacency matrix are non-zero, where we assume that the target Y has only incoming edges and the variables are arranged in topological order. For the two-branch graph, values in the adjacency are set to zero accordingly. For chain graphs, the first lower off-diagonal entries are non-zero. The exogenous variables $\boldsymbol{U}$ and shift interventions $\boldsymbol{i}$ are independent Gaussian with $U_j, i_j \sim \mathcal{N}(0, 1)$ for $j = 1, ..., N$.

Data and Training The data and training parameters are summarized in Table 1. All simulation data is generated before training and reused in each epoch. We split the data into training (70%), validation and test (15% each). Since the training of the two-variable model would occasionally get stuck in local minima, we run each training with 10 different random initializations of the weights and select the model with the best total validation loss (14) at the end of training. Furthermore, we use a cosine annealing learning rate scheduler with a final learning rate of 10^{-5}.

G.3 Experimental Details: Double Well

Simulation We model the ball moving in a double well potential $V(x) = x^4 - 4x^2$, shown in Figure 7(a), by the following equation of motion:

$$m\ddot{x}(t) + k\dot{x}(t) + \frac{\partial}{\partial x}V(x(t)) = 0 \quad \Rightarrow \quad m\ddot{x}(t) + k\dot{x}(t) + 4x(t)^3 - 8x(t) = 0\,, \tag{30}$$

where $x(t)$ is the position of the ball at time t, $\dot{x}(t)$ and $\ddot{x}(t)$ are the first and second time derivatives, respectively, k is the friction coefficient and m is the mass of the ball. We can reformulate the second

order ODE into a system of first order ODEs by introducing the velocity $v(t) = \dot{x}(t)$ as a variable:

$$\begin{aligned} \dot{x}(t) &= v(t) \\ \dot{v}(t) &= -\frac{1}{m}\left(k v(t) + 4x(t)^3 - 8x(t)\right) . \end{aligned} \tag{31}$$

We solve the system of ODEs numerically on a grid of 101 time points t_k for $k = 0, \ldots, 100$ equally spaced between $t = 0$ and $t = 10$ using a numerical integration method. The initial conditions are $x(0) = -2.07414285 + 5 \times 10^{-7} \times \varepsilon_x$, with $\varepsilon_x \sim \text{Uniform}(-1, 1)$ and $v(0) = 11$. The initial values are chosen such that there is a non-zero chance that the ball ends up in the left or right well without any additional interventions.

For shift interventions, we sample random velocity shifts $\Delta v(t_k) \sim \mathcal{N}(0, 0.5)$. The positions are unshifted. In the numerical integration scheme, the shift interventions are implemented by splitting the integration domain in parts. The ODE system is integrated from the initial conditions at t_0 to the next time grid at t_1. Then the velocity at t_1 is shifted by $\Delta v(t_1)$ and used as the initial value for the next integration starting at t_1, and so on. Similarly, we introduce independent stochasticity by adding noise to the velocity sampled from $\mathcal{N}(0, 0.2)$ at each time step, mimicking intrinsic noise of the system.

Parameters	**Double Well** (Fig. 7)
learning rate λ	$5 \cdot 10^{-4}$
learning rate scheduler	-
No. repeated train. runs per seed	1
simulation paths n_{sim}	$10,000$
training epochs N_{ite}	200
simulation batch size B	128
intervention batch size B_i	64
overlap reg. η_{ov} (12)	0
balancing reg. η_{bal} (13)	0

Table 2: **Experimental parameters and settings for the double well experiments.**

Data and Training The data and training parameters are summarized in Table 1. All simulation data is generated before training and reused in each epoch. We split the data into training (70%), validation and test (15% each).

G.4 Experimental Details: Spring-Mass System

Parameters	**4 masses with different weights** (Fig. 2)	**2 groups of masses** (Fig. 8)
learning rate λ	10^{-4}	10^{-3}
learning rate scheduler	cosine annealing	cosine annealing
No. repeated train. runs per seed	5	5
simulation paths n_{sim}	$10,000$	$10,000$
training epochs N_{ite}	$4,800$	$1,800$
simulation batch size B	128	128
intervention batch size B_i	64	512
overlap reg. η_{ov} (12)	0.1	0.1
balancing reg. η_{bal} (13)	0.1	0.1
spring constant k	10^{-3}	10^{-3}
rest length u_0	1	1
masses m_i	$(0.5, 0.83, 0.17, 1.5)$	all 1

Table 3: **Experimental parameters and settings for the spring mass system experiments.**

Simulation Let M be the number of masses. Then, $m_i \in \mathbb{R}$, $\vec{\tilde{x}}_i(t) \in \mathbb{R}^2$ and $\vec{v}_i(t) \in \mathbb{R}^2$ represent the weight, position and velocity of mass $i = 1, \ldots, M$ at time t. $A \in \{0,1\}^{M \times M}$ is the adjacency matrix encoding the spring connections, where $A_{ij} = 1$ indicates that a spring connects masses i and j. The rest length at which the springs exert no force is denoted by u_0 and k is the spring constant. Both u_0 and k are assumed to be the same for all springs.

The total force acting on mass i at time t is given by

$$\vec{F}_i(t) = -k \sum_{j, A_{ij}=1} \left(\|\vec{u}_{ij}(t)\| - u_0 \right) \frac{\vec{u}_{ij}(t)}{\|\vec{u}_{ij}(t)\|} \tag{32}$$

where $\vec{u}_{ij}(t) = \vec{x}_i(t) - \vec{x}_j(t)$ is the displacement vector from mass j to mass i. The equations of motion are

$$\frac{\mathrm{d}\vec{\tilde{x}}_i(t)}{\mathrm{d}t} = \vec{v}_i(t), \qquad \frac{\mathrm{d}\vec{v}_i(t)}{\mathrm{d}t} = \vec{a}_i(t), \quad \text{with} \quad \vec{a}_i(t) = \frac{\vec{F}_i(t)}{m_i}\,. \tag{33}$$

We assume that the masses have no volume and do not collide or interact other than the forces coming from the springs.

We solve the system of ODEs numerically on a grid of 21 time points t_k for $k = 0, \ldots, 20$ equally spaced between $t = 0$ and $t = 100$ using a numerical integration method. The positions are initially set on a grid to $\vec{\tilde{x}}_1(t=0) = (0,0) + \vec{\tilde{x}}_{\text{offset}}$, $\vec{\tilde{x}}_2(t=0) = (1,0) + \vec{\tilde{x}}_{\text{offset}}$, $\vec{\tilde{x}}_3(t=0) = (0,1) + \vec{\tilde{x}}_{\text{offset}}$ and $\vec{\tilde{x}}_4(t=0) = (1,1) + \vec{\tilde{x}}_{\text{offset}}$, where $\vec{\tilde{x}}_{\text{offset}} \sim \mathcal{N}(0, 10)$ is a random offset that shifts the entire system. The initial velocities are independently drawn as $\vec{v}_i(t=0) \sim \mathcal{N}(0, 0.01)$. We apply random independent velocity shifts $\Delta\vec{v}_i(t_k) \sim \mathcal{N}(0, 0.005)$ at each time step and integrate it into the ODE solver in the same way as for the double well experiment in App. G.3.

The feature vectors $\boldsymbol{X}$ used to learn the TCR of the spring-mass system consists of all velocity values for all masses across all simulated time points. The interventions $\boldsymbol{i}$ are the corresponding velocity interventions.

Data and Training The data and training parameters are summarized in Table 3. All simulation data is generated before training and reused in each epoch. We split the data into training (70%), validation and test (15% each). Similar to the experiments on the two-branch linear graph in App. G.2, we repeat the training runs with different weight initializations and use a cosine annealing learning rate scheduler.